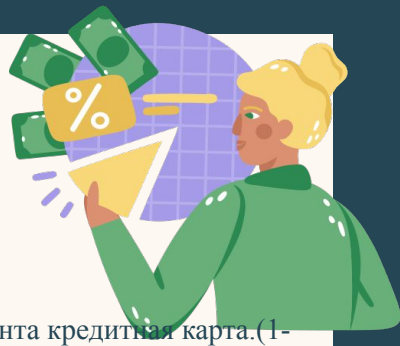
Прогнозирование оттока клиентов банка

Выполнила:
Коркина Елизавета Алексеевна

Dataset о клиентах, предоставленный ABC Multinational Bank

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Card Type	Point Earned
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1	1	2	DIAMOND	464
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	1	3	DIAMOND	456
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1	1	3	DIAMOND	377
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0	0	5	GOLD	350
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	0	5	GOLD	425
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0	0	1	DIAMOND	300
9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0	0	5	PLATINUM	771
9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1	1	3	SILVER	564
9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1	1	2	GOLD	339
10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0	0	3	DIAMOND	911

О наборе данных



RowNumber — соответствует номеру записи (строки).

CustomerId — ID клиента.

Surname— фамилия клиента.

CreditScore — кредитный рейтинг.

Geography — местоположение клиента.(Germany,France,Spain)

Gender -пол.

Age — возраст.

Tenure — обозначает количество лет, в течение которых клиент был клиентом банка.

Balance — баланс на счетах.

NumOfProducts — количество продуктов, которые клиент приобрел через банк.

HasCrCard — указывает, есть ли у клиента кредитная карта.(1-да,0-нет)

IsActiveMember — активные клиенты.(1-да,0-нет)

EstimatedSalary — ориентировочная заработная плата.

Exited — покинул ли клиент банк.(1-да,0-нет)

Complain — есть у клиента жалоба или нет.(1-да,0-нет)

Satisfaction Score — оценка, предоставленная клиентом за решение жалобы.(1-5)

Card Type — тип карты, которую держит клиент.(DIAMOND,GOLD,SILVER,PLATINUM)

Point Earned — баллы, заработанные клиентом за использование кредитной карты.

Card Type:
- Diamond
- Gold
- Silver
- Platinum

Geography:
- France
- Germany
- Spain

Gender:
- Male
- Female

Dataset содержит 10000 строк и 15 столбцов

Object переменными являются Geography, Gender, Card Type

Данные не содержат нулевые значения, так как автор предоставил их уже очищенными

Цель проекта

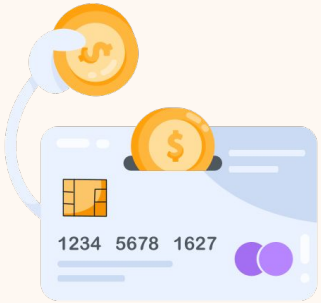
Целью проекта является создание модели, которая поможет прогнозировать отток клиентов банка.

Целевой переменной будет являться Exited, которая принимает значение 1,если клиент покинул банк и 0-не покинул.

Актуальность

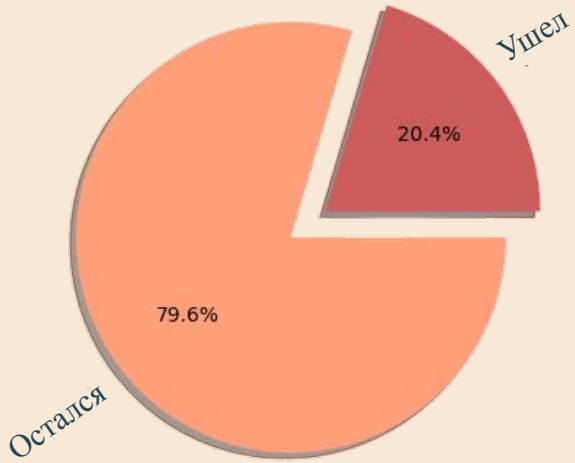
Проект по прогнозированию оттока клиентов банка является весьма актуальным и имеет большое практическое значение для финансовых учреждений. В современной конкурентной банковской среде удержание клиентов является одной из ключевых задач для обеспечения стабильности и роста бизнеса.Отток клиентов, то есть потеря клиентской базы, может оказывать негативное влияние на финансовые показатели банка, такие как доли рынка, прибыльность, рентабельность.
Кроме того, привлечение новых клиентов обычно требует больших затрат, поэтому сохранение существующих клиентов является более эффективным с точки зрения будзатрат.

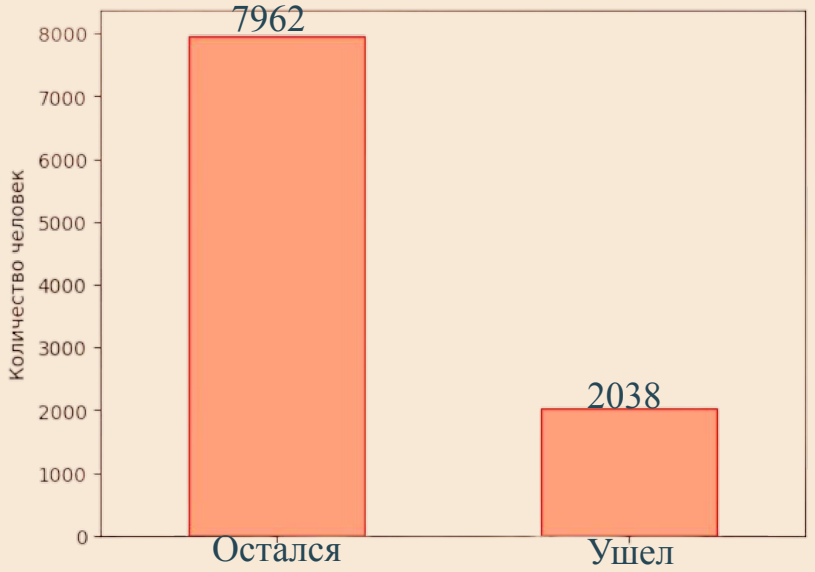
Задачи проекта



1. Выбор данных.Описание датасета,цели прогнозирования.
2.Провести разведочный анализ данных(EDA).
3.Предобработать данные для дальнейшего обучения модели.
4.Выбрать модель для прогнозирования оттока клиентов банка.
5.Разделить данные на обучающую и тестовую выборку. Обучающая выборка будет использоваться для обучения модели, а тестовая выборка позволит оценить производительность модели на новых данных.
6.Обучить модель на обучающей выборке и оценить ее производительность на тестовой выборке, используя метрики.

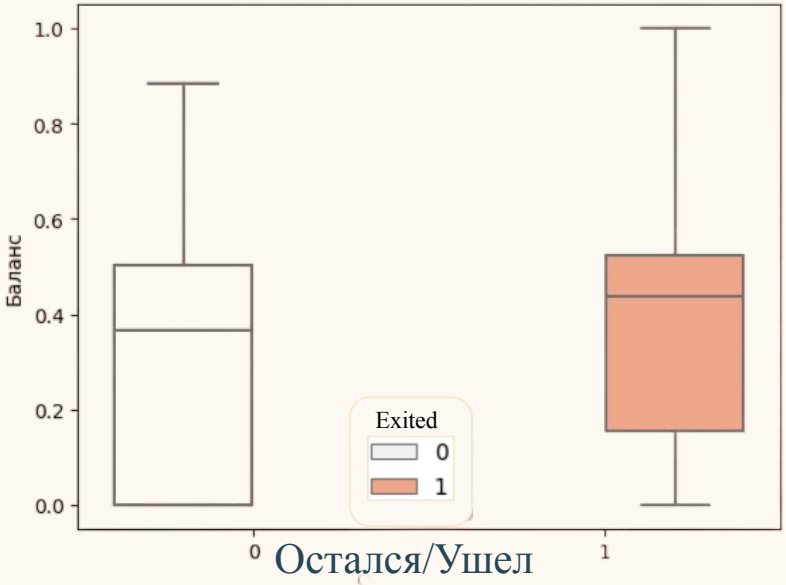
Целевая переменная Ушел/Остался клиент



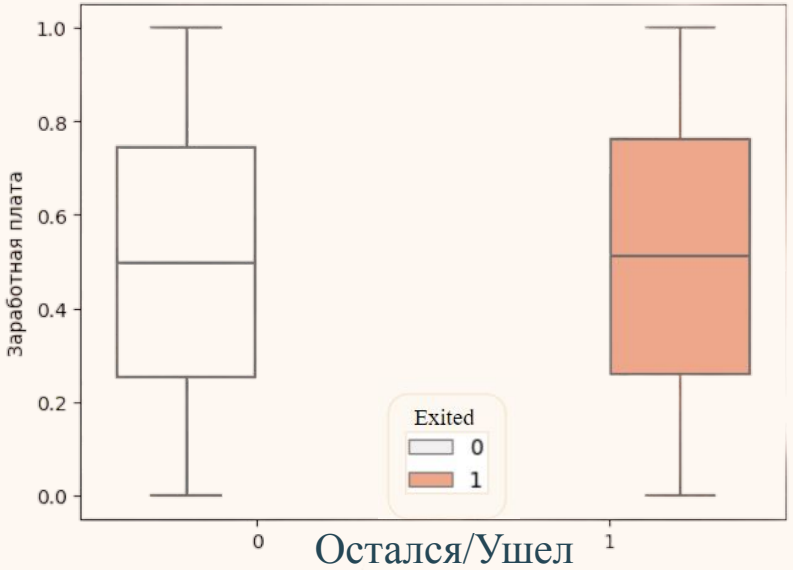


Рассмотрим интересные зависимости

Баланс на карте-Ушел/Остался клиент

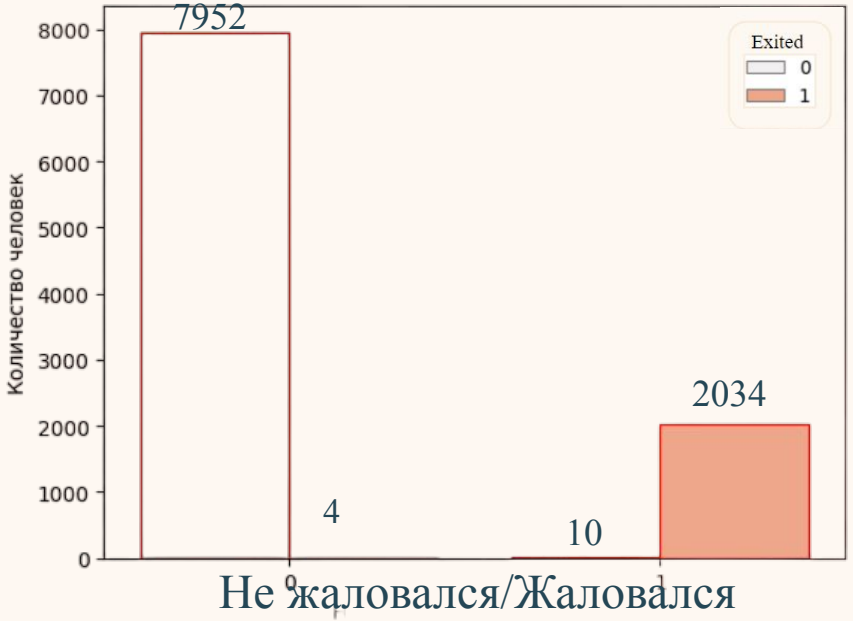


Заработная плата-Ушел/Остался клиент

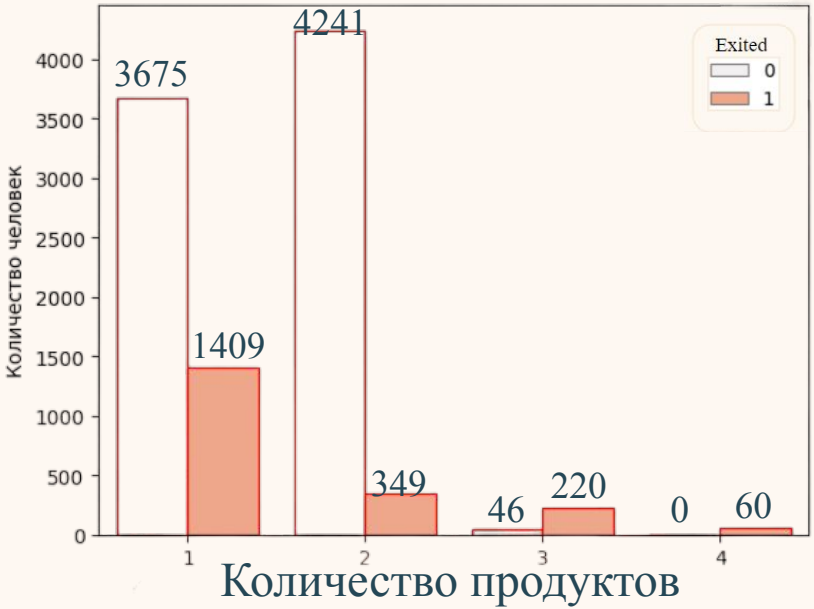


Рассмотрим интересные зависимости

Жаловался/Не жаловался-Ушел/Остался клиент

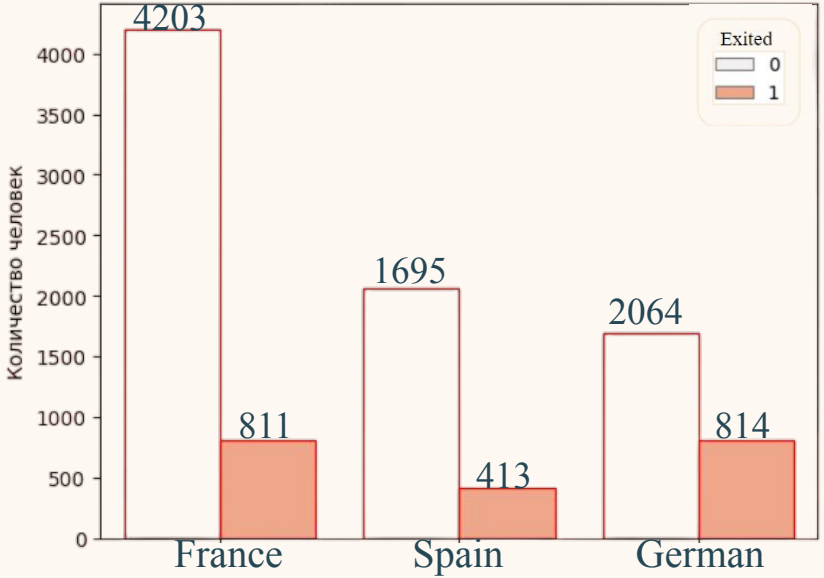


Количество продуктов, приобретенных через банк-Ушел/Остался клиент

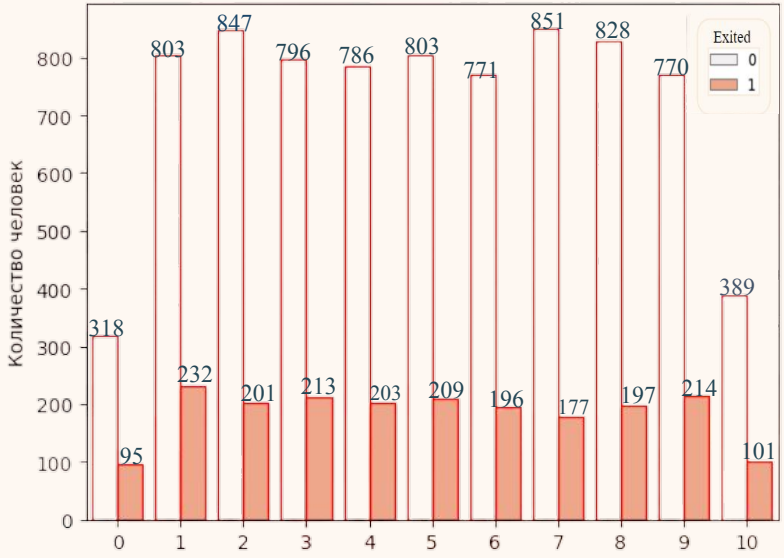


Рассмотрим интересные зависимости

Местоположение -Ушел/Остался клиент



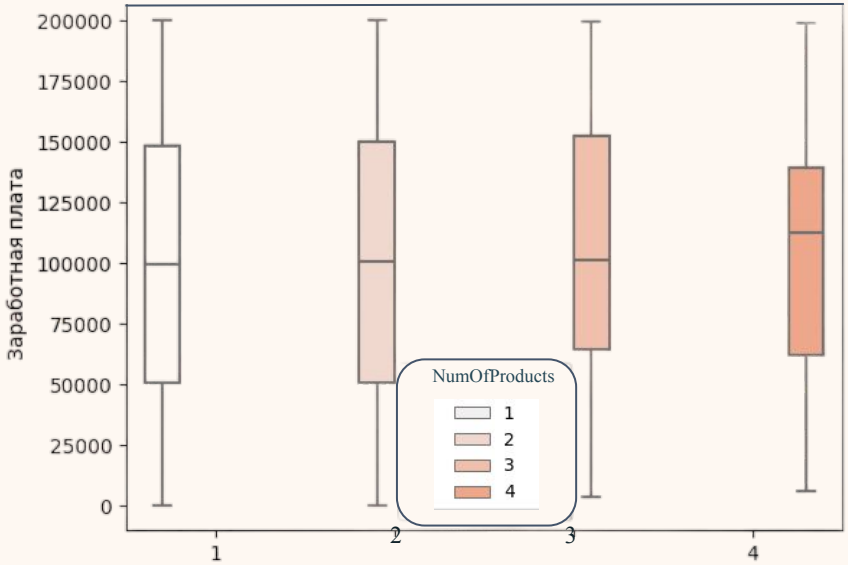
Количество лет,в течение которых клиент был клиентом банка-Ушел/Остался клиент



Количество лет, в течение которых клиент был клиентом банка

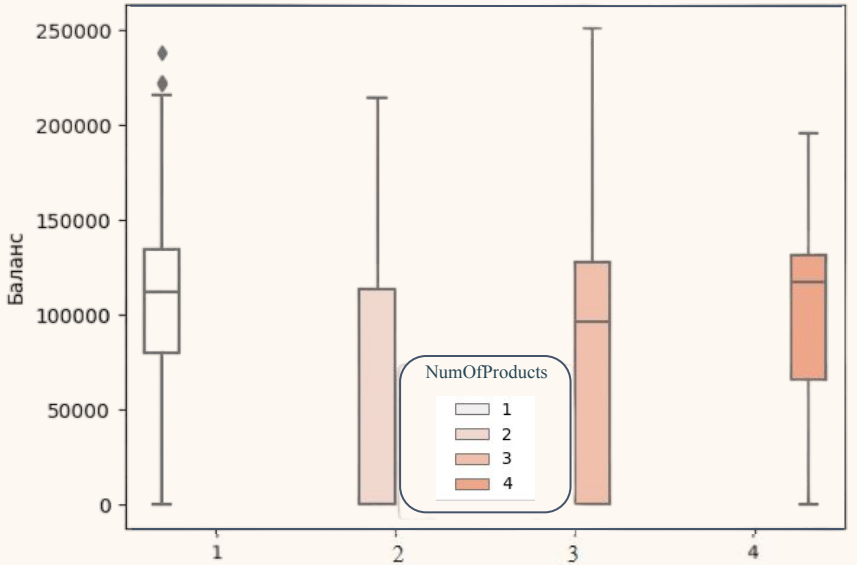
Рассмотрим интересные зависимости

Количество продуктов,приобретенных через банк-Заработная плата клиента



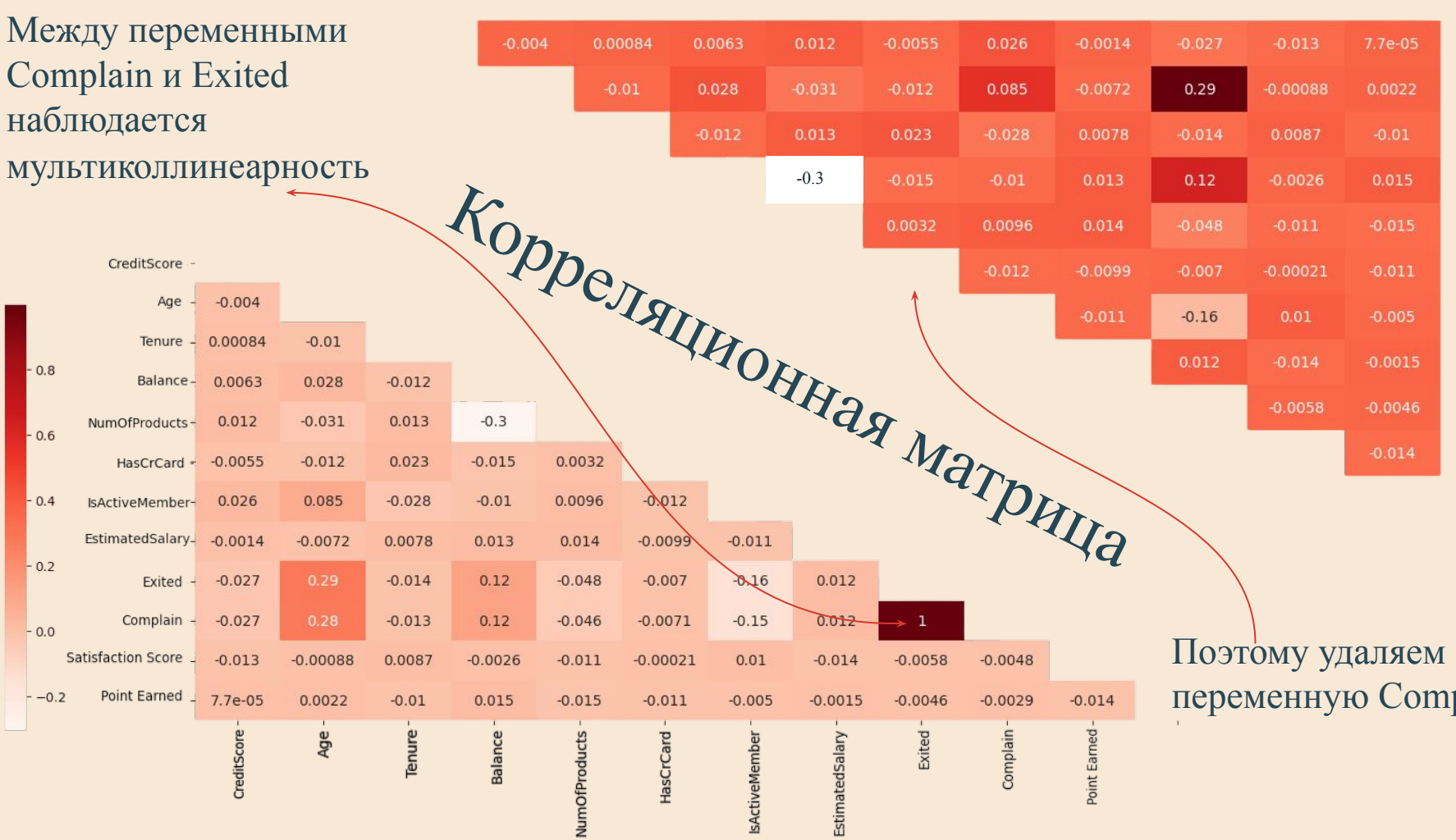
Количество продуктов, приобретенных через банк

Количество продуктов,приобретенных через банк-Баланс на карте клиента



Количество продуктов, приобретенных через банк

Между переменными Complain и Exited наблюдается мультиколлинеарность

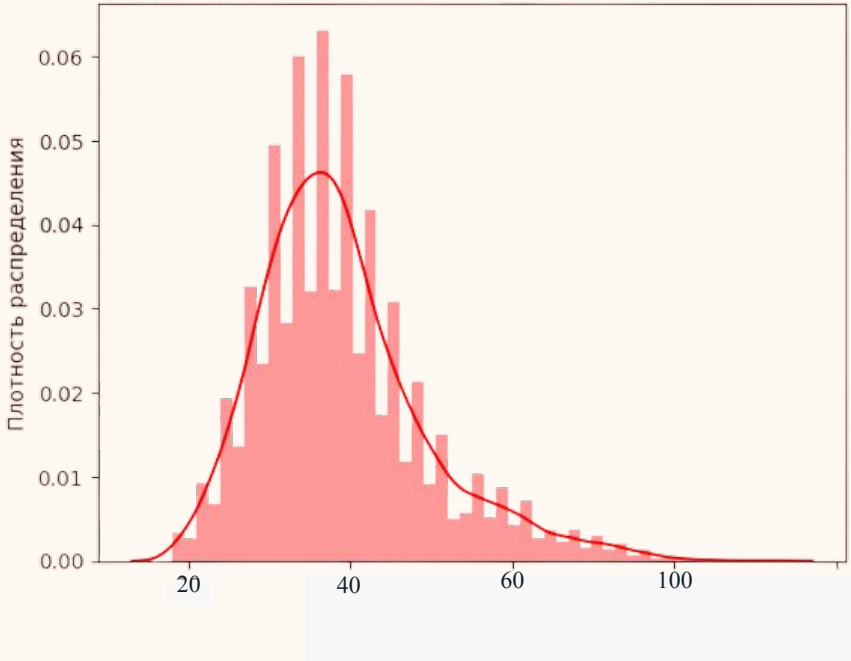


Поэтому удаляем переменную Complain

Аномалии

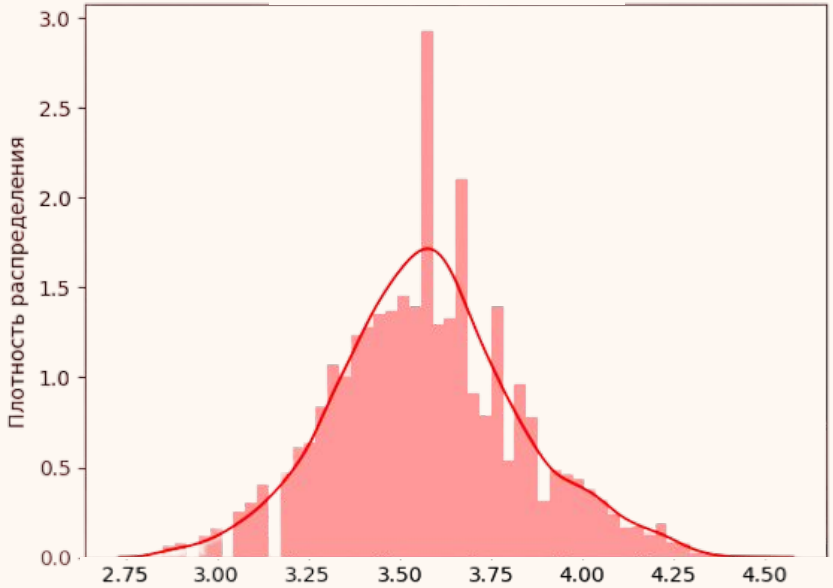
Плотность распределения с выбросами

Возраст



Плотность распределения без выбросов

Возраст



Масштабируем непрерывные переменные при помощи метода MinMax

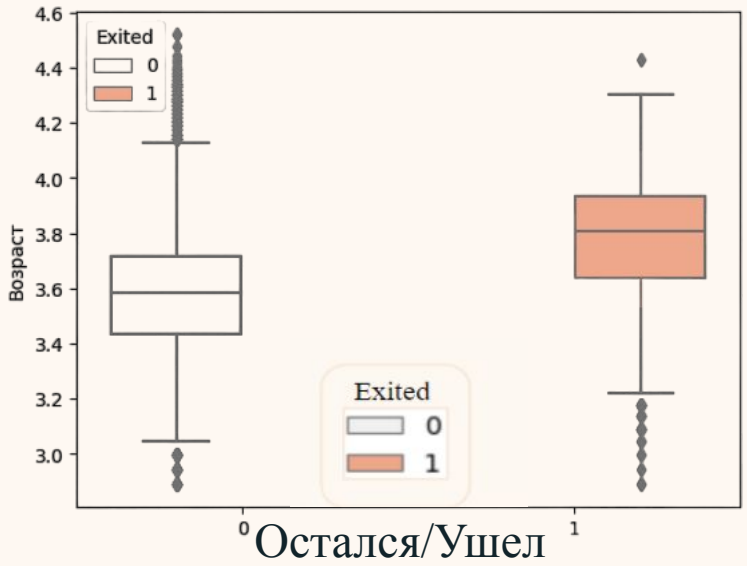
Перепишем переменную Пол:Female=1,Male=0

Сделаем из категориальных переменных дамми переменные

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Point Earned	Geography_France	Geography_Germany	Geography_Spain	Card Type_DIAMOND	Card Type_GOLD	Ty
0	0.538	1	0.519363	2	0.000000	1	1	1	0.506735	1	1	2	0.391600	1	0	0	1	0	
1	0.516	1	0.504592	1	0.334031	1	0	1	0.562709	0	1	3	0.382520	0	0	1	1	0	
2	0.304	1	0.519363	8	0.636357	3	1	0	0.569654	1	1	3	0.292849	1	0	0	1	0	
3	0.698	1	0.473938	1	0.000000	2	0	0	0.469120	0	0	5	0.262202	1	0	0	0	1	
4	1.000	1	0.533787	2	0.500246	1	1	1	0.395400	0	0	5	0.347333	0	0	1	0	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9995	0.842	0	0.473938	5	0.000000	2	1	0	0.481341	0	0	1	0.205448	1	0	0	1	0	
9996	0.332	0	0.407607	10	0.228657	1	1	1	0.508490	0	0	5	0.740068	1	0	0	0	0	
9997	0.718	1	0.424874	7	0.000000	1	0	1	0.210390	1	1	3	0.505108	1	0	0	0	0	
9998	0.844	0	0.519363	3	0.299226	2	1	0	0.464429	1	1	2	0.249716	0	1	0	0	1	
9999	0.884	1	0.270828	4	0.518708	1	1	0	0.190914	0	0	3	0.898978	1	0	0	1	0	

10000 rows × 20 columns

Проверим гипотезы



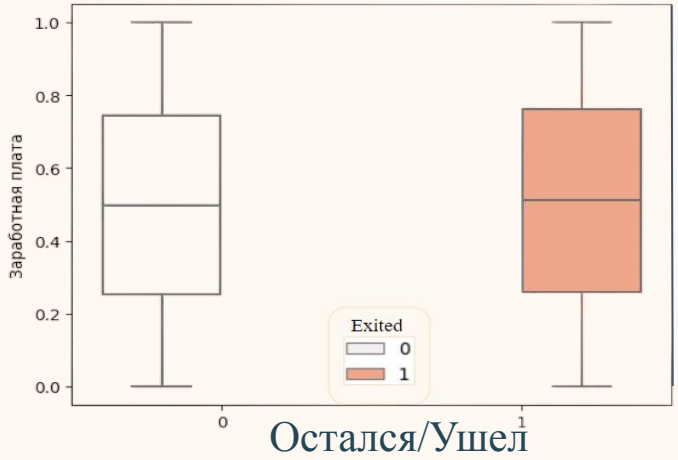
Средний возраст ушедших клиентов равен среднему возрасту оставшихся клиентов

$$H_0 : \mu_{\text{Возраст ушедших клиентов}} = \mu_{\text{Возраст оставшихся клиентов}}$$

$$H_1 : \mu_{\text{Возраст ушедших клиентов}} > \mu_{\text{Возраст оставшихся клиентов}}$$

Вывод:
Наблюдаемое значение 32.5 больше критического значения 1.64, значит нулевая гипотеза отвергается
Таким образом средний возраст ушедших клиентов выше, чем средний возраст клиентов оставшихся в банке.

Проверим гипотезы



Средняя заработная плата ушедших клиентов равна средней заработной плате оставшихся клиентов

$$H_0 : \mu_{\text{Заработная ушедших клиентов}} = \mu_{\text{Заработная оставшихся клиентов}}$$

$$H_1 : \mu_{\text{Заработная ушедших клиентов}} > \mu_{\text{Заработная оставшихся клиентов}}$$

Вывод:
Наблюдаемое значение 1.24 меньше критического 1.64,значит нулевая гипотеза не отвергается.Таким образом средняя заработная плата ушедших клиентов равна средней заработной плате оставшихся клиентов.

Модели

- Метод K-ближайших соседей
- Логистическая регрессия
- Случайный лес
- Метод опорных векторов
- Решающие деревья
- Градиентный бустинг

Обучающая выборка: все признаки, кроме целевой переменной Exited(80%)

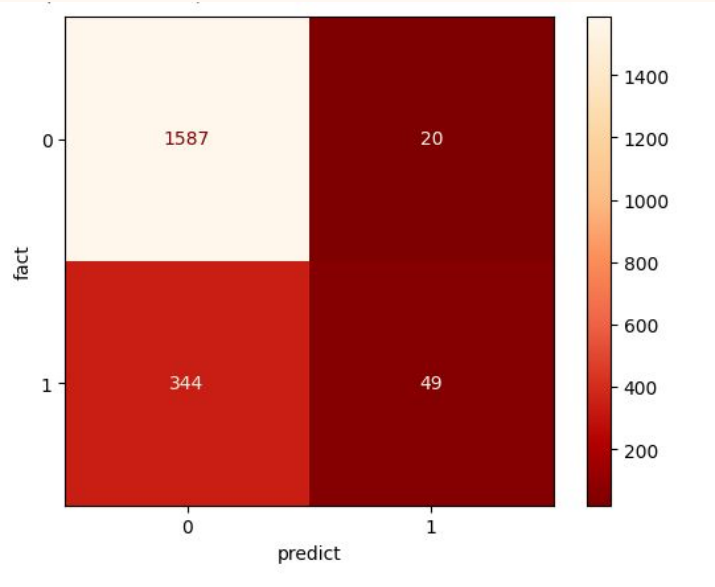
Тестовая выборка: Целевая переменная Exited(20%)

Так как тестовая выборка несбалансированная, то для сравнения моделей будем использовать метрику recall,так как нужно минимизировать количество объектов класса 1, которые модель ошибочно отнесла к классу 0(по предсказанию модели клиент останется, по факту клиент уйдет)

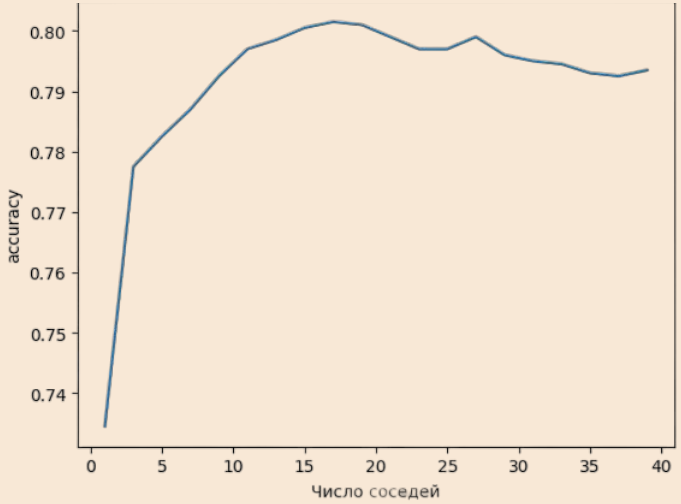
Метод K-ближайших соседей

Выбранные гиперпараметры:

```
{'n_neighbors': 19, 'p': 3, 'weights': 'uniform'}
```



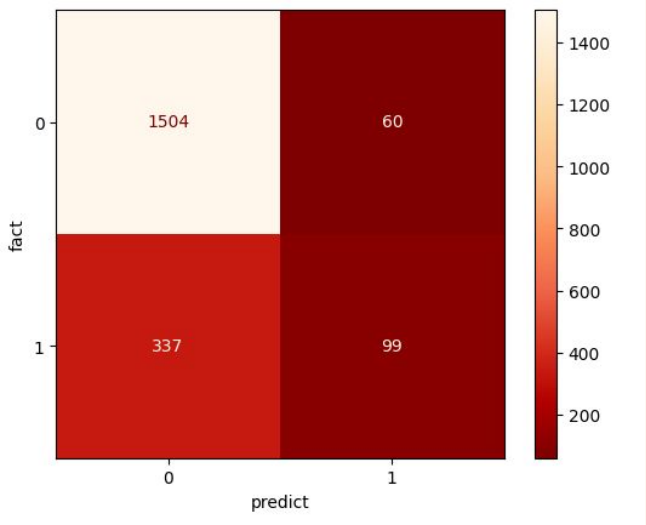
	precision	recall	f1-score	support
0	0.81	0.98	0.89	1587
1	0.62	0.10	0.17	413
accuracy			0.80	2000
macro avg	0.71	0.54	0.53	2000
weighted avg	0.77	0.80	0.74	2000

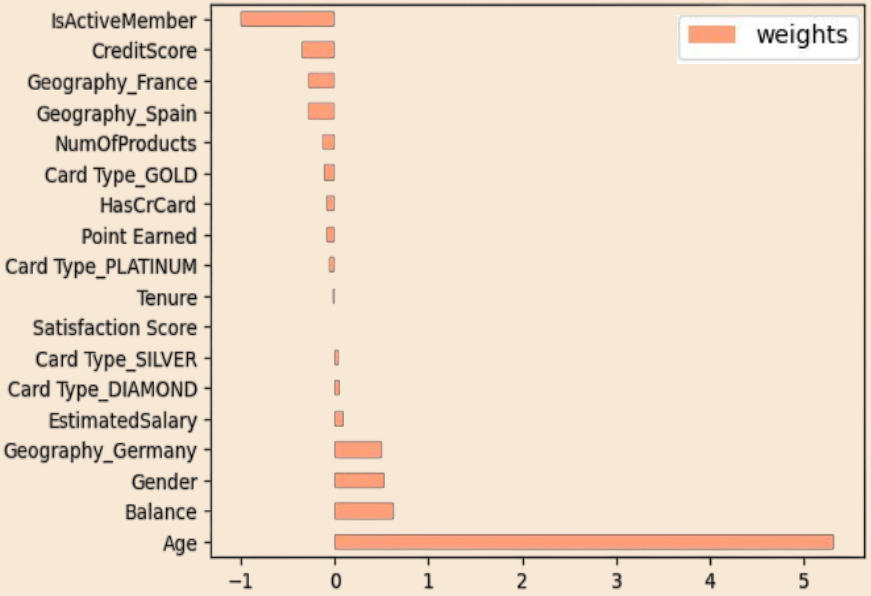


Логистическая регрессия

Выбранные гиперпараметры:

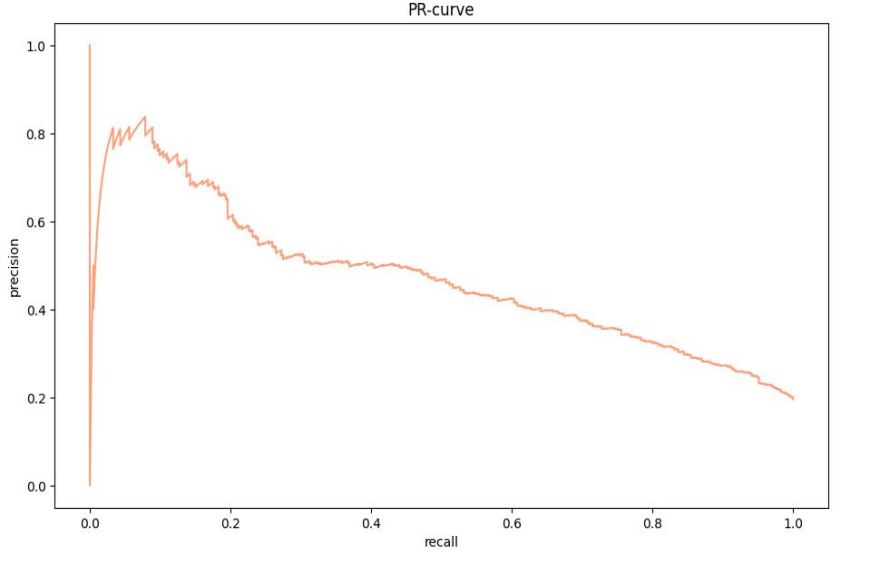
```
{'C': 215.44346900318823, 'penalty': 'l2'}
```

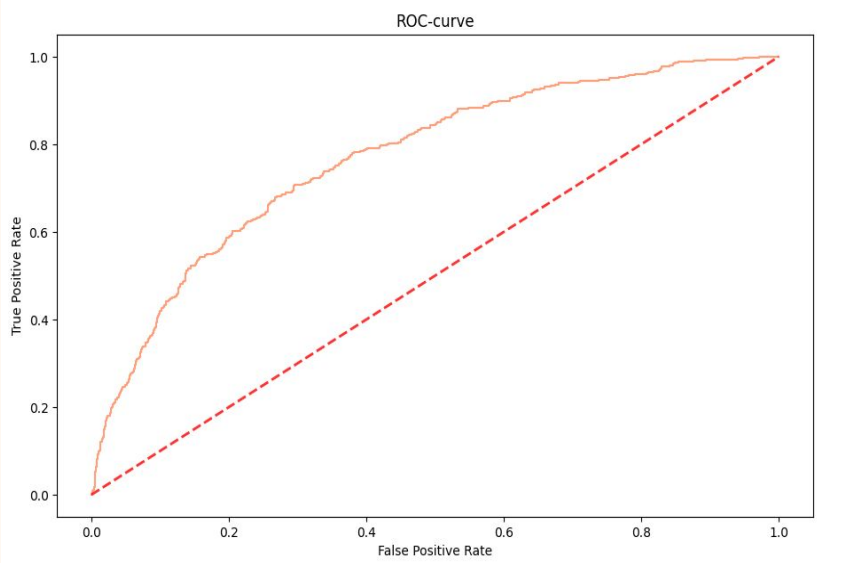




	precision	recall	f1-score	support
0	0.82	0.96	0.88	1567
1	0.61	0.23	0.33	433
accuracy			0.80	2000
macro avg	0.71	0.59	0.61	2000
weighted avg	0.77	0.80	0.76	2000

Логистическая регрессия

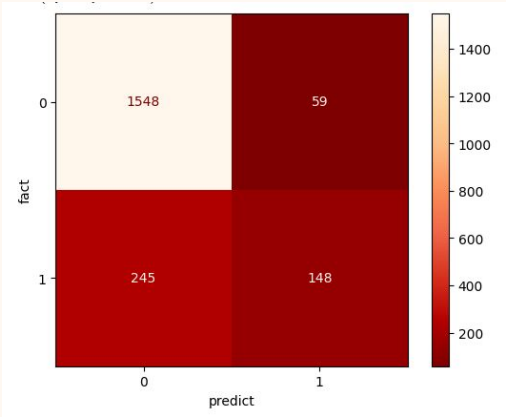


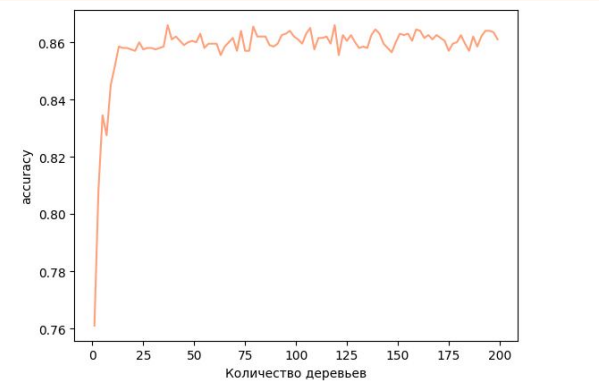


AUC-PR=0.47

AUC-ROC=0.78

Случайный лес

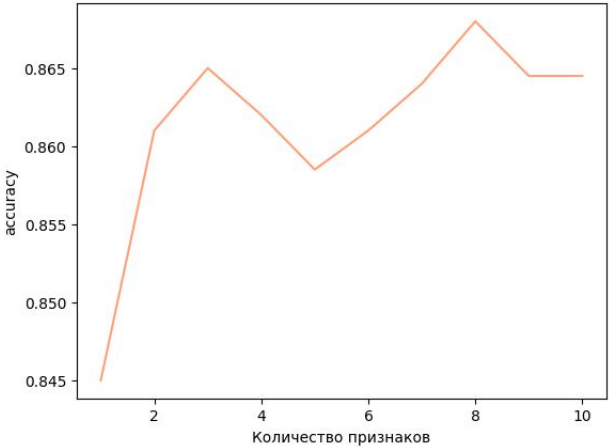




Выбранные гиперпараметры:

```
{'max_depth': 4, 'max_features': 8, 'n_estimators': 150}
```

	precision	recall	f1-score	support
0	0.82	1.00	0.90	1567
1	0.97	0.21	0.35	433
accuracy			0.83	2000
macro avg	0.89	0.60	0.62	2000
weighted avg	0.85	0.83	0.78	2000

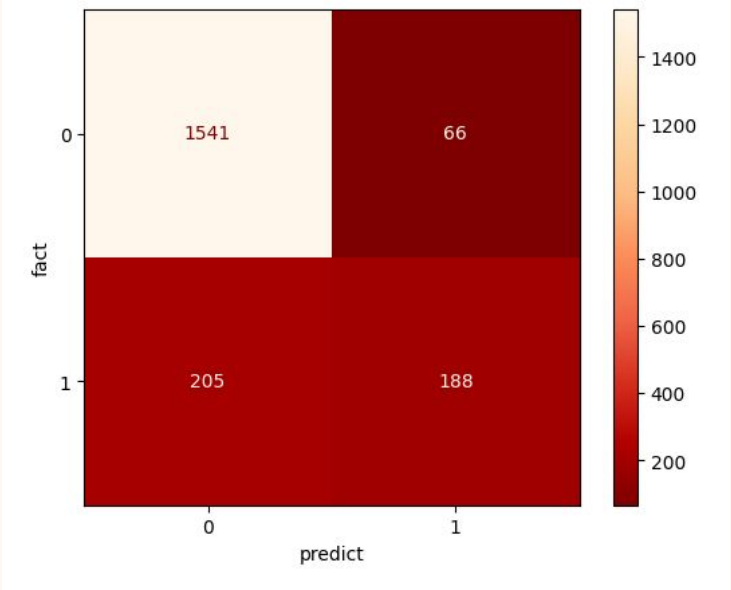


{'max_depth': 4, 'max_features': 8, 'n_estimators': 150}

Метод опорных векторов

Выбранные гиперпараметры:

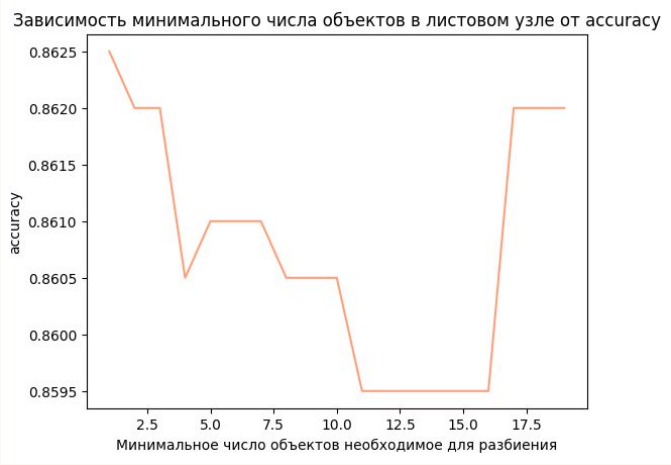
```
{'C': 100, 'kernel': 'rbf'}
```



	precision	recall	f1-score	support
0	0.86	0.98	0.92	1588
1	0.84	0.40	0.54	412
accuracy			0.86	2000
macro avg	0.85	0.69	0.73	2000
weighted avg	0.86	0.86	0.84	2000

Решающее дерево

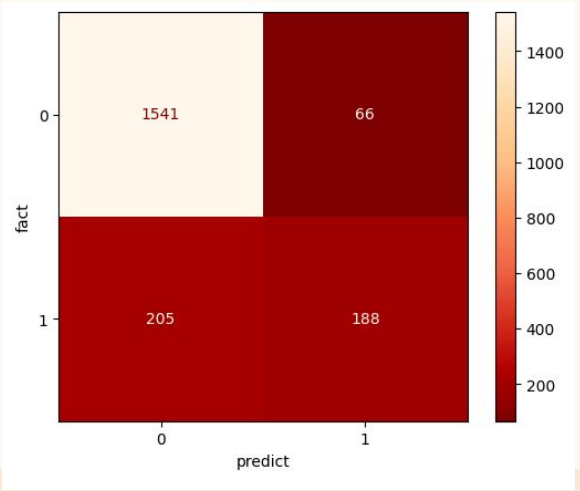
	max_depth	accuracy
0	16	0.7990
1	19	0.8000
2	17	0.8035
3	18	0.8050
4	15	0.8115
5	14	0.8145
6	2	0.8180
7	13	0.8205
8	11	0.8235
9	12	0.8305
10	3	0.8370
11	10	0.8390
12	9	0.8470
13	4	0.8485
14	5	0.8555
15	6	0.8625
16	8	0.8645
17	7	0.8650



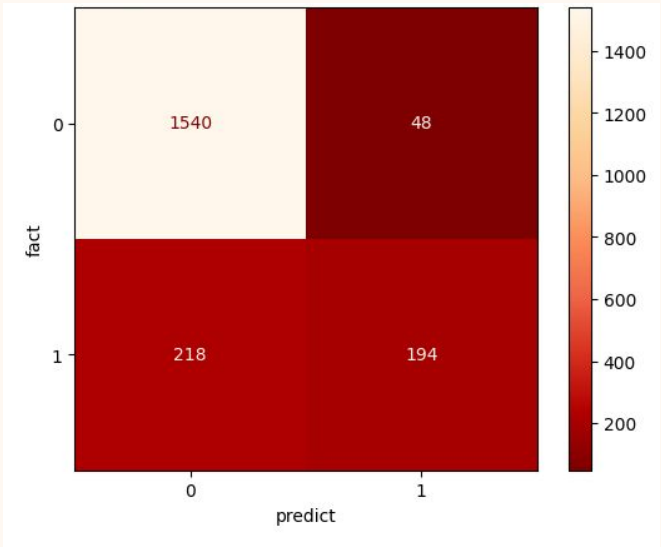
Выбранные гиперпараметры:

```
{'ccp_alpha': 0.001,
 'criterion': 'entropy',
 'max_depth': 8,
 'max_features': 'auto'}
```

	precision	recall	f1-score	support
0	0.85	0.95	0.90	1588
1	0.65	0.35	0.45	412
accuracy			0.83	2000
macro avg	0.75	0.65	0.67	2000
weighted avg	0.81	0.83	0.81	2000



Градиентный бустинг



Выбранные гиперпараметры:

```
{'learning_rate': 0.01, 'max_depth': 4, 'n_estimators': 500, 'subsample': 0.5}
```

	precision	recall	f1-score	support
0	0.88	0.97	0.92	1588
1	0.80	0.47	0.59	412
accuracy			0.87	2000
macro avg	0.84	0.72	0.76	2000
weighted avg	0.86	0.87	0.85	2000

Вывод

Лучшей моделью будет Градиентный бустинг, так как метрика recall больше у этой модели

	Модели	recall
0	К-ближайших соседей	0.114679
1	Логистическая регрессия	0.227064
2	Случайный лес	0.029817
3	Метод опорных векторов	0.399083
4	Градиентный бустинг	0.493119
5	Дерево решений	0.110092

	0_pred	1_pred
0_true	TN	FP
1_true	FN	TP