МИНОБРНАУКИ РОССИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Санкт-Петербургский политехнический университет Петра Великого»
Институт дополнительного образования
Высшая инженерная школа

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

«Разработка модели предсказания оттока клиентов в сфере телекоммуникационных услуг»

по программе профессиональной переподготовки: «Анализ данных на языке Python»

Выполнила: Ефименко М.С.
Руководитель: к.т.н., Семендяев Р.Ю.

Санкт-Петербург
2025

Область исследования

Сфера телекоммуникаций отличается высокой конкуренцией, быстрым развитием технологий и сложными предпочтениями клиентов, что требует постоянного анализа поведения потребителей и разработки стратегий для их удержания.

Цель работы

Разработать модель предсказания оттока клиентов в телекоммуникационной сфере, используя исторические данные о клиентах и их взаимодействии с услугами компании.

Задачи:

- Провести предварительный анализ данных, построить визуализации, проверить баланс целевой переменной;
- Исследовать ключевые характеристики клиентов;
- Выбор модели для предсказания;
- Подбор гиперпараметров. Выбор лучшей модели;
- Оценка модели;
- Разделение клиентов на группы риска.

Данные для исследования

Набор данных включает данные о:

Клиентах, которые ушли в течение последнего месяца.

Услугах, на которые подписался каждый клиент — телефон, интернет, онлайн-безопасность, резервное копирование, техническая поддержка и потоковое телевидение и фильмы.

Информация об учетной записи клиента — как долго они являются клиентом, тип контракта, способ оплаты.

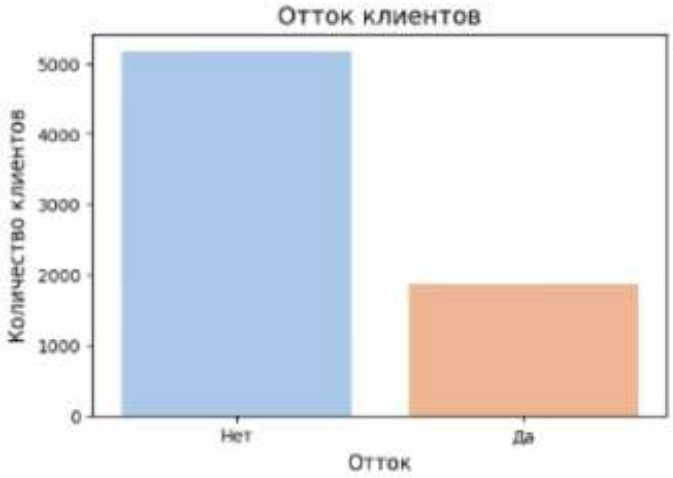
Демографическая информация о клиентах — пол, возрастной диапазон, а также есть ли у них партнеры и иждивенцы.

- Источник данных:

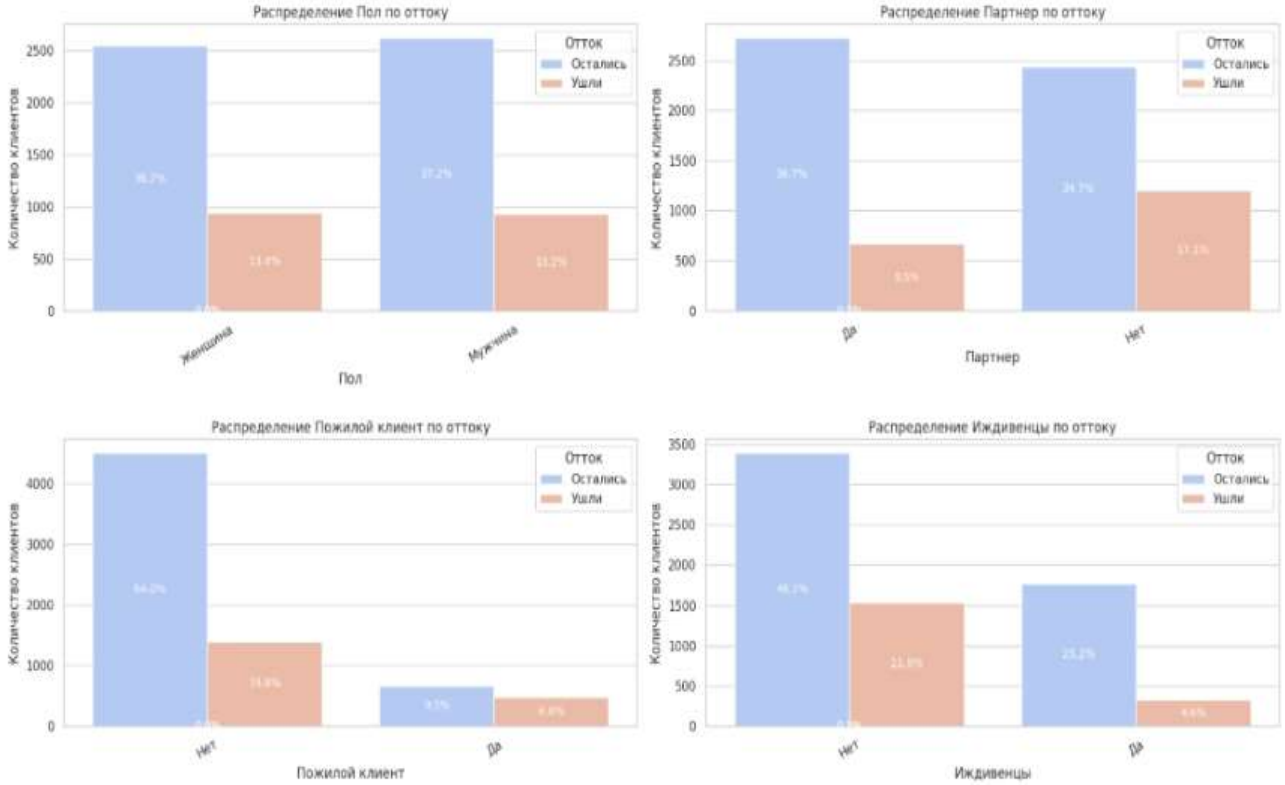
https://www.kaggle.com/datasets/blastchar/telco-customer-churn?select=WA_Fn-UseC_-Telco-Customer-Churn.csv

Первичный анализ данных

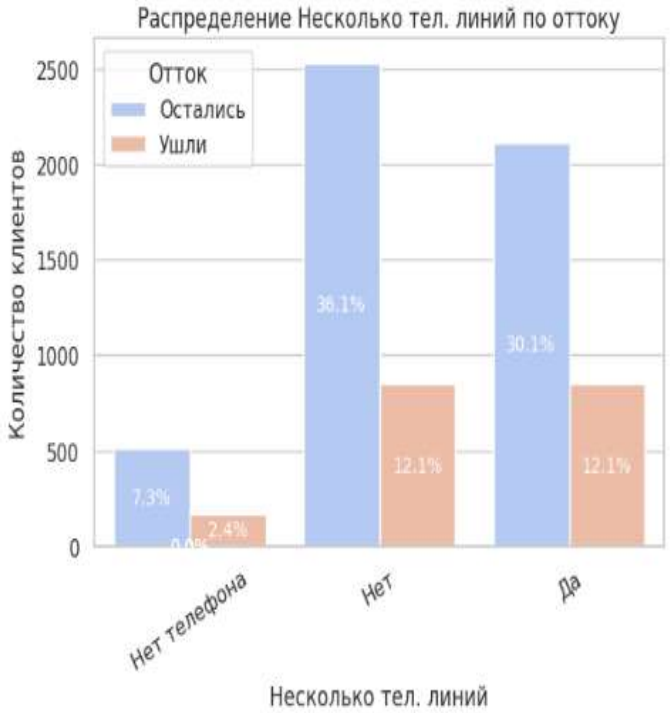
- Исходные данные состоят из 7043 строк и 20-ти столбцов, из которых 16 с факторными признаками и 1 результативный (целевой) столбец;
- Целевая переменная распределена неравномерно;
- Данные не имеют пропущенных значений;
- Факторные признаки в основном представлены категориальными данными.



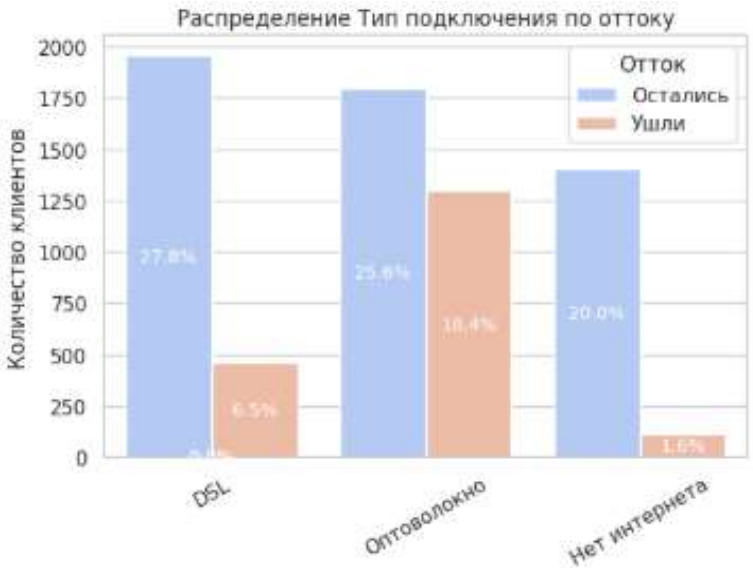
Влияние демографических признаков на отток клиентов

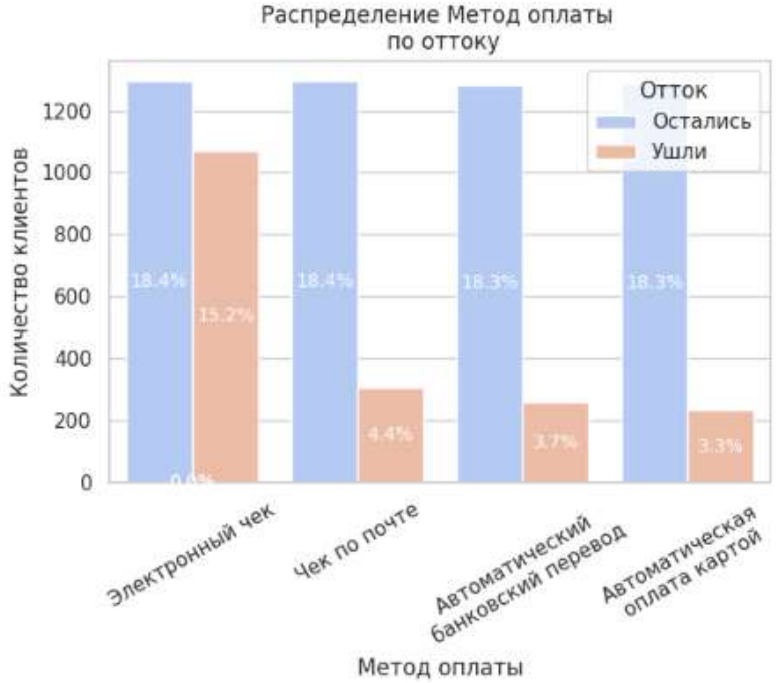


Влияние телефонной связи и способа оплаты на отток клиентов

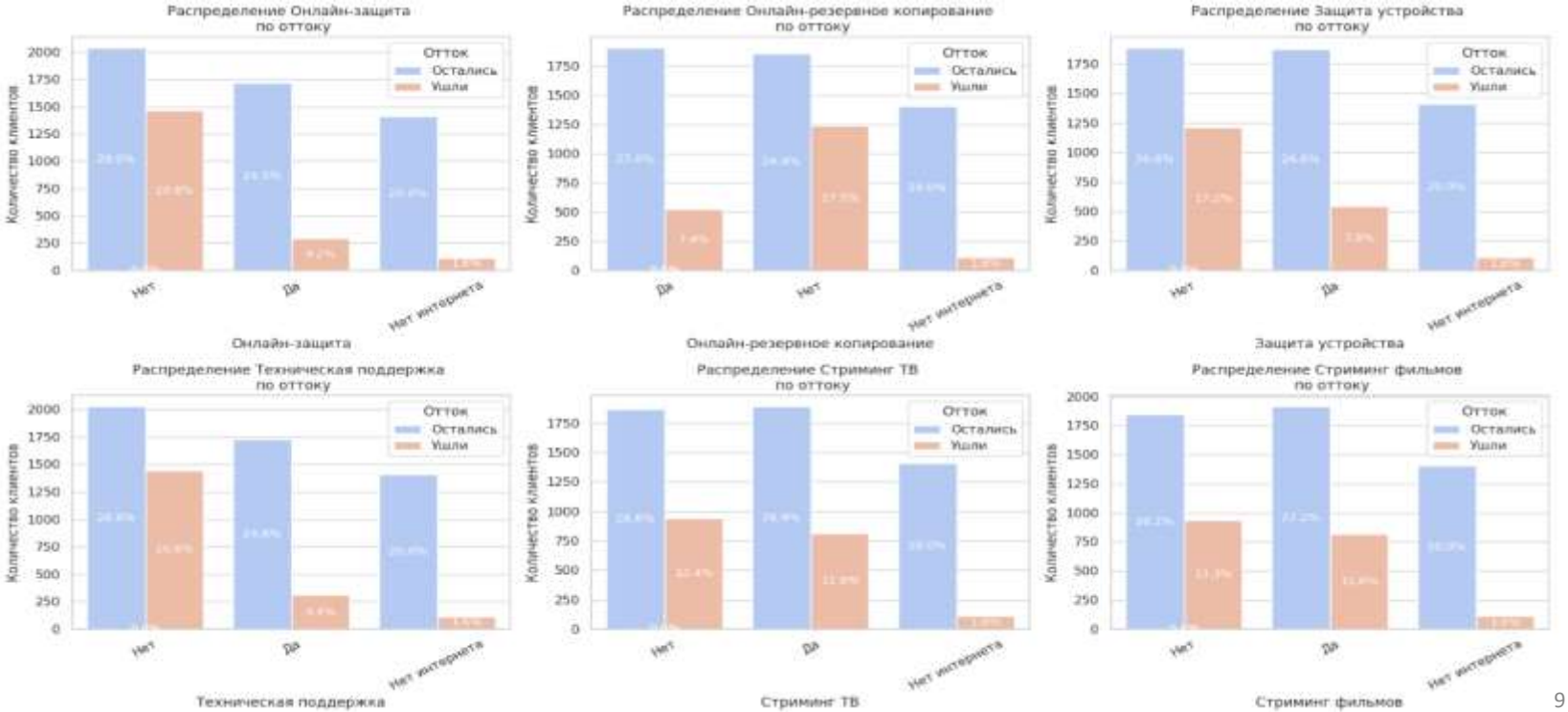


Зависимость оттока клиентов от типа интернет-подключения и способа оплаты

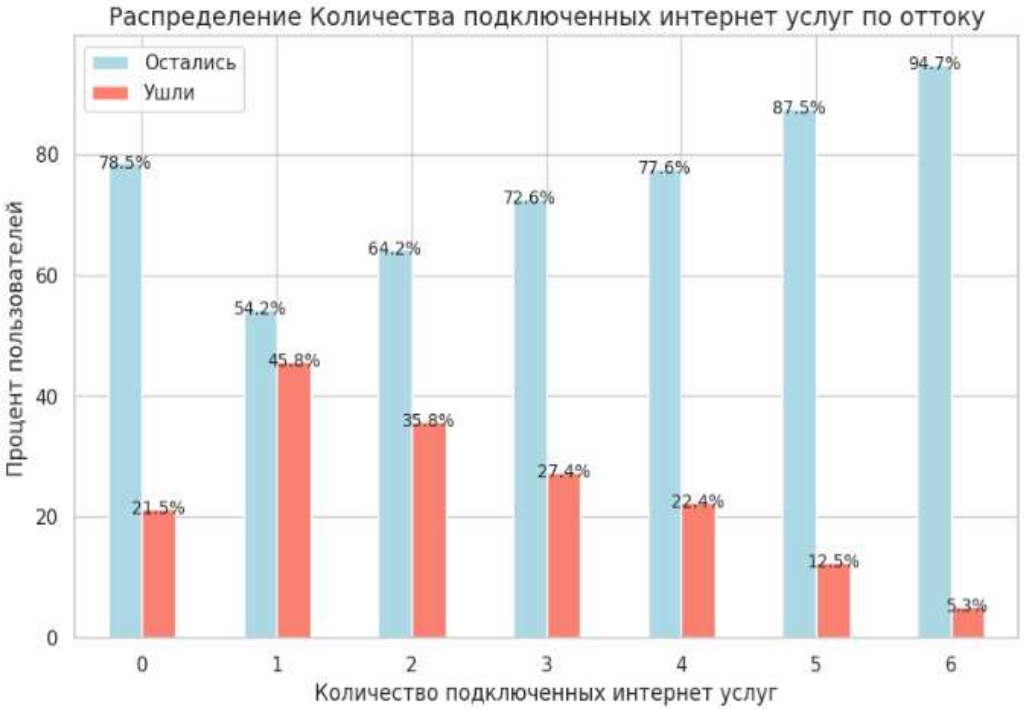




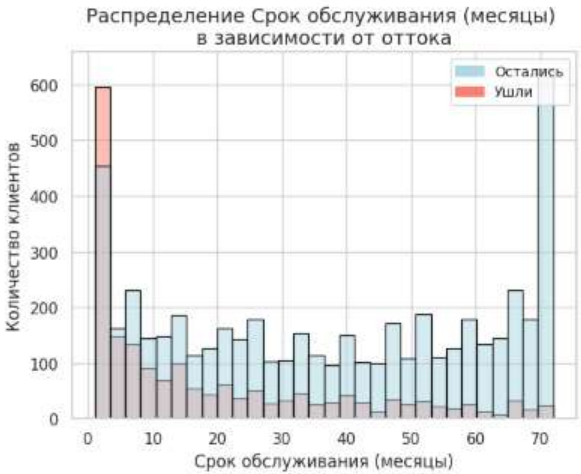
Связь между подключенными онлайн-услугами и оттоком клиентов

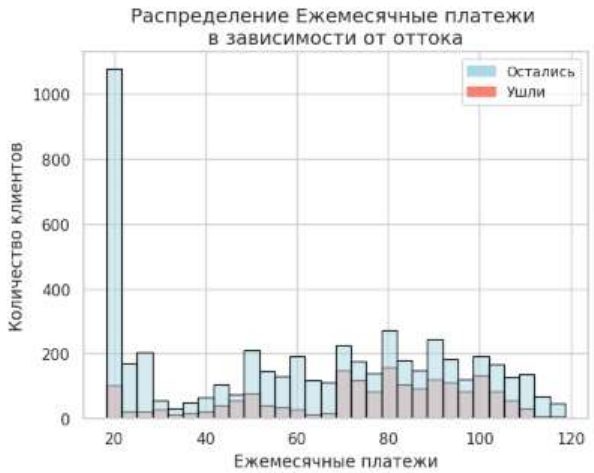


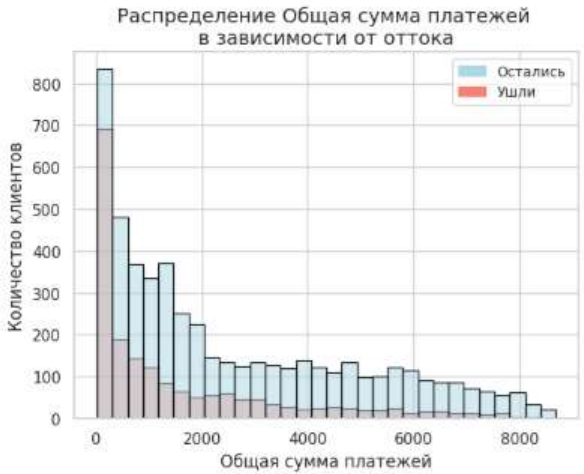
Влияние количества подключенных интернет-услуг на отток клиентов



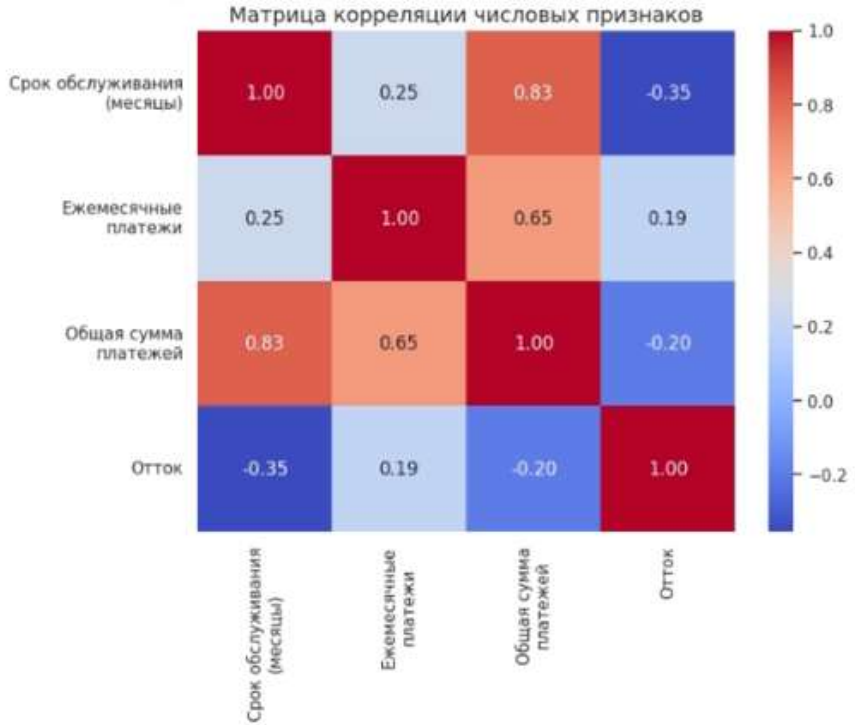
Влияние срока обслуживания, ежемесячных платежей и общей суммы платежей на отток клиентов







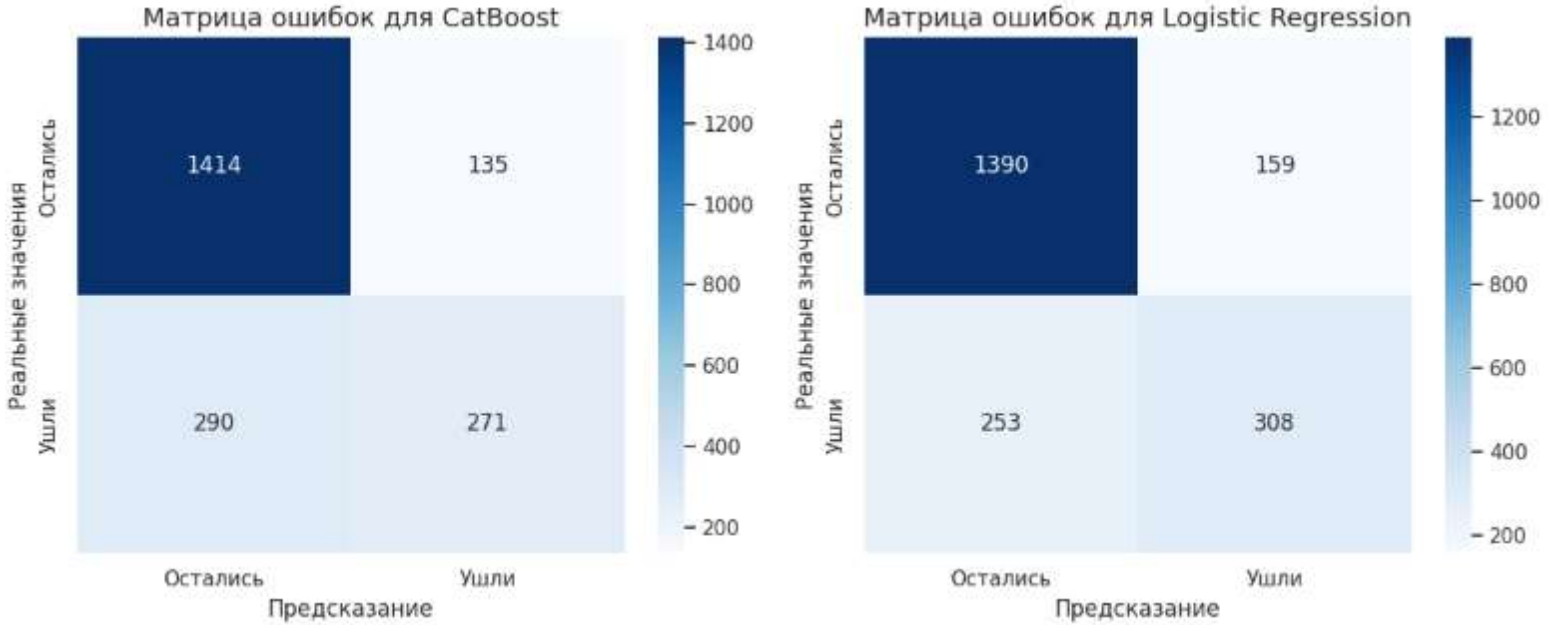
Анализ взаимосвязей между числовыми признаками



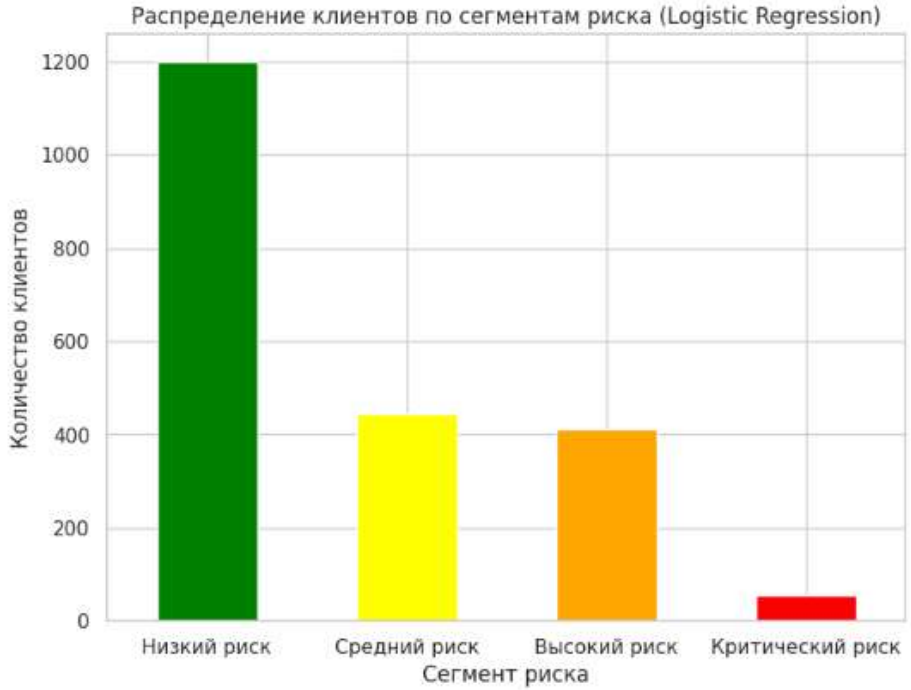
Сравнение метрик моделей до и после подбора гиперпараметров

	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.80427	0.79586	0.80427	0.79812	0.83837
Random Forest	0.78626	0.77115	0.78626	0.77054	0.82506
XGBoost	0.7763	0.76708	0.7763	0.77037	0.81375
CatBoost	0.79858	0.78637	0.7986	0.78631	0.8407
Random Forest (opt.)	0.7360	0.7617	0.7360	0.6269	0.8311
XGBoost (opt.)	0.7905	0.7792	0.7905	0.7817	0.8362
CatBoost(opt.)	0.7986	0.7867	0.7986	0.7872	0.8401

Матрицы ошибок для моделей Logistic Regression и CatBoost



Распределение клиентов по сегментам риска



Вывод:

- Проведен предварительный анализ и визуализация данных;
- Исследованы ключевые характеристики клиентов;
- Проведено сравнение моделей и отбор лучшей;
- Выполнена сегментация клиентов по рискам.

Логистическая регрессия оказалась наиболее эффективной моделью для прогнозирования оттока клиентов, так как показала наилучшие результаты.

Градиентные бустинги с подбором гиперпараметров (CatBoost, XGBoost) также показали хорошие результаты, но не превзошли логистическую регрессию.