Построение рекомендательной системы оттока клиентов для принятия маркетинговых решений

Дарья Мягкова

07.07.2023

Содержание

1 Введение и постановка задачи

Оператор связи «Ниединогоразрыва.ком» хочет научиться прогнозировать отток клиентов. Если выяснится, что пользователь планирует уйти, ему будут предложены промокоды и специальные условия. Команда оператора собрала персональные данные о некоторых клиентах, информацию об их тарифах и договорах.

Оператор предоставляет услуги стационарной телефонной связи и интернета. Также доступно подключение дополнительных услуг.

Решается задача классификации, где целевым будет бинарный признак факта ухода клиента из компании: 0 соответствует активному клиенту, 1 - ушедшему. Распределение значений обусловлено используемыми метриками: выбор и оптимизация моделей производится через призму максимизации метрики ROC-AUC, также учитывается время обучения модели. В качества дополнительной будет рассмотрена f1-метрика. Важно также обратить внимание на ложноотрицательные ответы, поскольку цель модели - идентиифкация клиента, планирующего смену оператора.

Планируется рассмотреть линейные модели (логистическая регрессия, Ridge), случайный лес, а также бустинги (градиентный бустинг, CatBoost). Обучение каждой проводится на обучающей выборке с кроссвалидацией и подбором гиперпараметров. Для наиболее оптимальной конфигурации итоговой модели вычисляется финальное значение целевой метрики на тестовой выборке. В соответствии с ТЗ, минимальный порог соответствует 0.85.

2 Описание данных

Сырые данные представляют из себя четыре датасета:

- contract_new.csv — информация о договоре;
- personal_new.csv — персональные данные клиента;
- internet_new.csv — информация об интернет-услугах;
- phone_new.csv — информация об услугах телефонии.

Во всех файлах столбец customerID содержит код клиента.
Информация о договорах актуальна на 1 февраля 2020.

Подключение интернета может быть двух типов: через телефонную линию (DSL) или оптоволоконный кабель (Fiber optic).

Также доступны такие услуги:

- Интернет-безопасность: антивирус (DeviceProtection) и блокировка небезопасных сайтов (OnlineSecurity);
- Выделенная линия технической поддержки (TechSupport);
- Облачное хранилище файлов для резервного копирования данных (OnlineBackup);
- Стриминговое телевидение (StreamingTV) и каталог фильмов (StreamingMovies).

Описание столбцов:

- BeginDate – дата начала пользования услугами;
- EndDate – дата окончания пользования услугами;
- Type – тип договора: ежемесячный, годовой и т.д.;
- PaperlessBilling – факт выставления счёта на электронную почту;
- PaymentMethod – способ оплаты;
- MonthlyCharges – ежемесячные траты на услуги;
- TotalCharges – всего потрачено денег на услуги;
- Dependents – наличие иждивенцев,
- Senior Citizen – наличие пенсионного статуса по возрасту;
- Partner – наличие супруга(и);
- MultipleLines – наличие возможности ведения параллельных линий во время звонка.

3 Чтение и предварительный обзор данных

Первый шаг - чтение датасетов. Для удобства дальнейшего объединения таблиц при формировании датасета с полным набором признаков в роли индекса выступает уникальный ID клиента.

Проведён предварительный обзор каждого датасета и изучение распределений значений каждого признака, в том числе на поиск пропусков, аномалий и тд.

3.1 Персональные данные клиента

Содержат информацию о каждом пользователе. Вид продемонстрирован на Рис. 1

customerID	gender	SeniorCitizen	Partner	Dependents
7590-VHVEG	Female	0	Yes	No
5575-GNVDE	Male	0	No	No
3668-QPYBK	Male	0	No	No
7795-CFOCW	Male	0	No	No
9237-HQITU	Female	0	No	No

Рис. 1: Обзор таблицы с данными о пользователях

7043 строки, 4 столбца. Явные пропуски отсутствуют. Тип данных у признака наличия пенсионного статуса - числовой, у остальных - строковый, все признаки категориальные. На рис. 2 продемонстрированы распределения значений в каждом столбце.

Можно выделить следующие особенности:

- количество женщин и мужчин в выборке примерно одинаковое
- 15 процентов выборки имеют пенсионный статус
- немного менее половины клиентов имеют партнёров
- две трети клиентов имеют иждивенцев

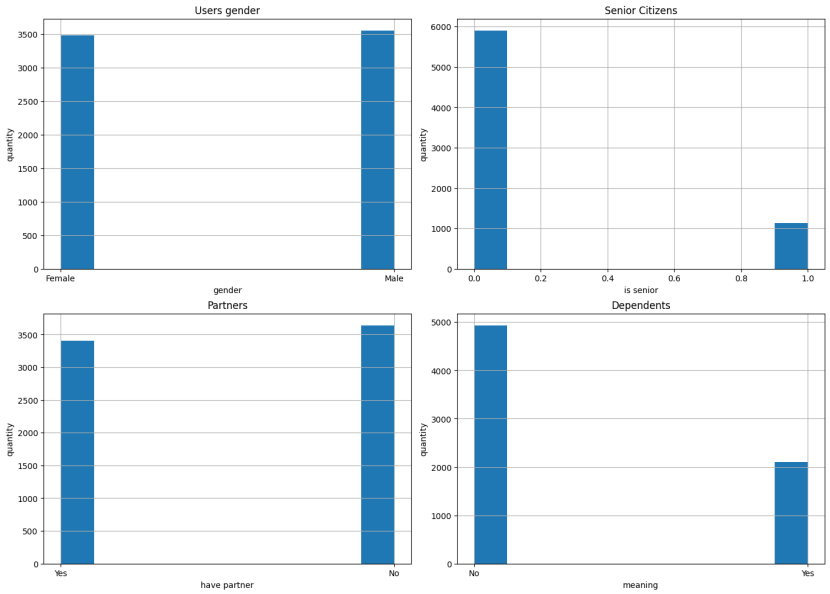

Рис. 2: Диаграмма моментов на участке выбора момента прокатки

3.2 Информация о договорах

Содержит 7043 строки (что соответствует количеству клиентов) и 7 признаков. Превью представлено на рис. 3 Явные пропуски отсутствуют. Тип данных - строковый для всех признаков кроме ежемесячных расходов.

customerID	BeginDate	EndDate	Type	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges
7590-VHVEG	2020-01-01	No	Month-to-month	Yes	Electronic check	29.85	31.04
5575-GNVDE	2017-04-01	No	One year	No	Mailed check	56.95	2071.84
3668-QPYBK	2019-10-01	No	Month-to-month	Yes	Mailed check	53.85	226.17
7795-CFOCW	2016-05-01	No	One year	No	Bank transfer (automatic)	42.30	1960.6
9237-HQITU	2019-09-01	No	Month-to-month	Yes	Electronic check	70.70	353.5

Рис. 3: Предварительный обзор датасета с характеристиками заключенных договоров

Обнаружено явное несоответствие типов:

– даты начала и окончания контракта: ожидается тип данных datetime, а здесь - строки: нюанс выгрузки данных, возможно - наличие неявных пропусков и тд

– сумма общих расходов за всё время: ожидается числовой формат вместо строкового.

Сначала изучим представленные на рис.. 4 распределения для характеристик , в которых тип данных соответствует ожидаемому:

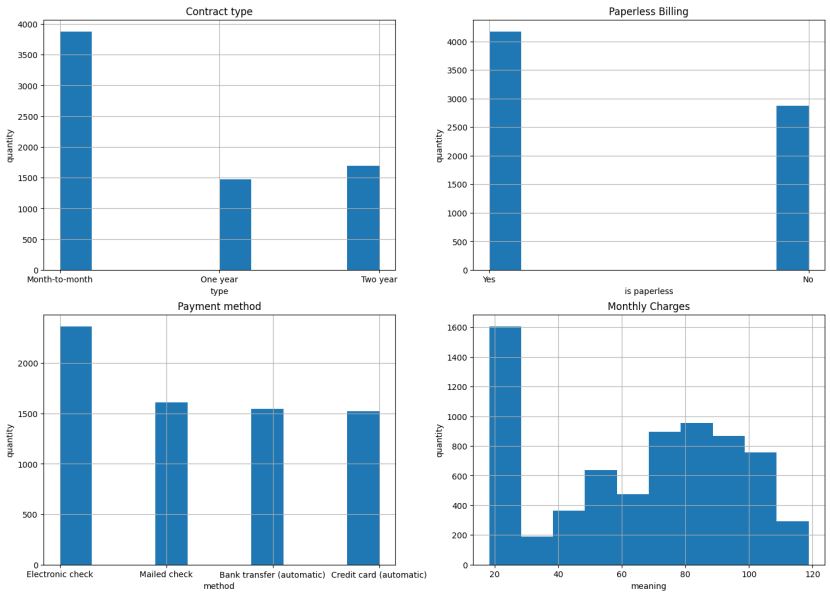

Рис. 4: Гистограмма распределений значений в характеристиках договора

Выделены следующие особенности в распределениях признаков:

- более половины клиентов предпочитают помесячные платежи. Оставшаяся часть распределена примерно поровну между годовыми и двухлетними контрактами, но последних немного больше
- порядка шестидесяти процентиов клиентов предпочитают оплату с чеком на электронную почту
- среди способов оплаты лидирует электронный чек. Чек по почте, автоматический банковский перевод и автосписание с карты примерно одинаково популярны по количеству пользователей
- распределение по месячным расходам имеет два пика. Первый, явно выраженный - в районе двадцати долларов. Вероятно, это со-

ответсвует ежемесячному платежу по минимальному тарифу. Если отбросить этот пик - остальная часть распределения похожа на нормальное с пиком в районе 80-90 долларов - возможно, это соответствует усреднённому платежу клиентов премиум-сегмента, пользующихся расширенным тарифным пакетом.

При переводе признака общего количества расходов из строкового типа данных в числовой обнаружено наличие неявных пропусков. Они соответствуют новым клиентам, подписавшим договор в феврале 20-ого года. То есть они ещё не успели внести первый платёж, или же этот платёж ещё не отображался в базе на момент выгрузки - соответственно, пропуски заполнены нулевым значением, и затем построено распределение, представленное на рис. 5.

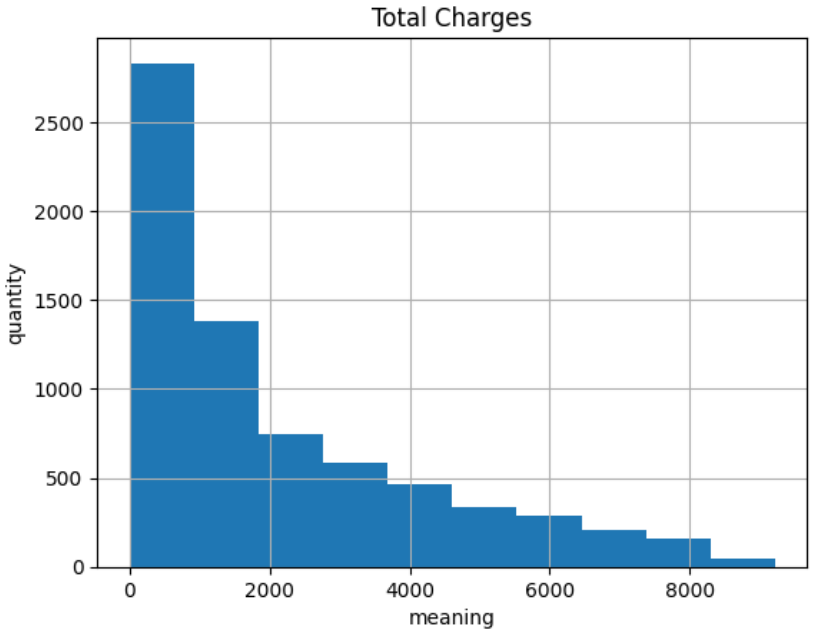

Рис. 5: Гистограмма распределения общих расходов пользователей

Распределение признака суммарного количества расходов имеет вид распределения Пуассона с пиком в районе тысячи долларов. Большинство клиентов сосредоточено в интервале от 0 до 2 тысяч долларов.

Далее на рис. 6 представлено распределение даты начала договора с точностью до года.

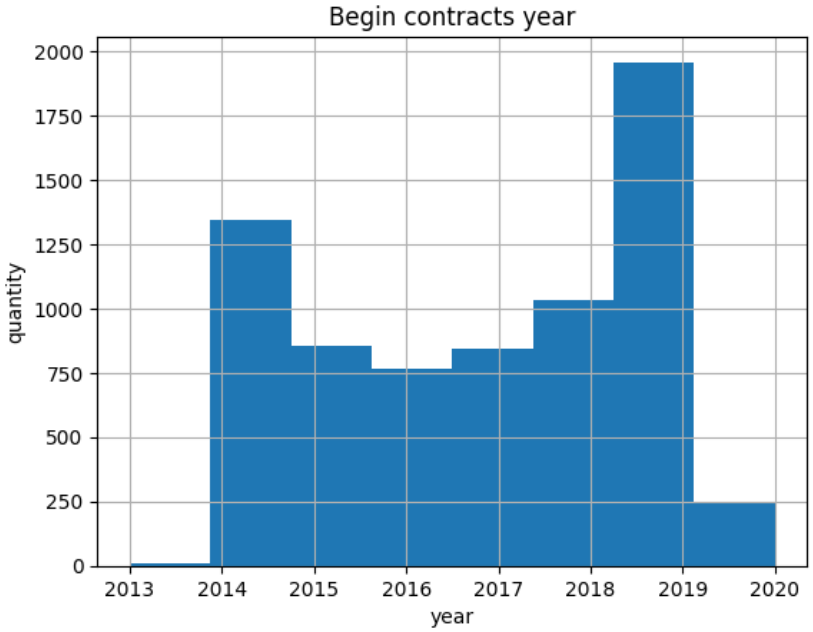

Рис. 6: Гистограмма распределения года подписания договора

Разброс - с 2013-ого до 2020-ого года выгрузки. Клиентов с 2013 года очень мало - либо вопрос репрезентативности выборки, либо компания открылась в 2013ом, и клиентов было совсем немного Спад в 2015ом году после 2014ого может быть обусловлен валютным кризисом. После локального пика в 2014-ом году наблюдается спад до 2016, и далее ежегодный прирост клиентов возрастает. В 2019-ом заметен резкий скачок: количество новых клиентов возросло в 2 раза по сравнению с 2018-ым годом. В 2020ом году договоров мало - база была выгружена в первой трети года.

Проведён предварительный обзор даты окончания договора. В нём два варианта значений - "No"и дата в строчном формате, соответствующие активному и ушедшему клиентам соответственно. Установлено, что количество незаконченных контрактов составляет 5942 штуки, что соответствует 15 процентам данных. Это значит, что присутствует существенный дисбаланс целевого признака.

3.3 Интернет - услуги

Таблица, представленная на рис. 7 состоит из 5517 строк и 7 столбцов.

customerID	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
7590-VHVEG	DSL	No	Yes	No	No	No	No
5575-GNVDE	DSL	Yes	No	Yes	No	No	No
3668-QPYBK	DSL	Yes	Yes	No	No	No	No
7795-CFOCW	DSL	Yes	No	Yes	Yes	No	No
9237-HQITU	Fiber optic	No	No	No	No	No	No

Рис. 7: Обзор датасета с информацией об интернет-услугах

Все признаки - категориальные, со строковым типом данных. Явных пропусков нет. Описывает тип подключения и набор логических идентификатор подключения услуг из пула возможных.

В распределениях признаков, представленных на рис. 8 , можно отметить следующие закономерности:

- подключение через оптоволоконный кабель на 8 процентов популярнее подулючения через телефонную линию DSL
- пользователей с включённой услугой блокировки небезопасных сайтов треть
- подключение облачного хранилища: количество пользователей примерно равно, но тех, у которых данная услуга не подключена, чуть больше
- подключение антивируса: количество пользователей в этих категориях примерно равно, но тех, в кото подключена данная услуга, меньше
- техподдержкой пользуется треть пользователей
- стриминговое телевидение и каталог фильмов подключены у половины пользователей.

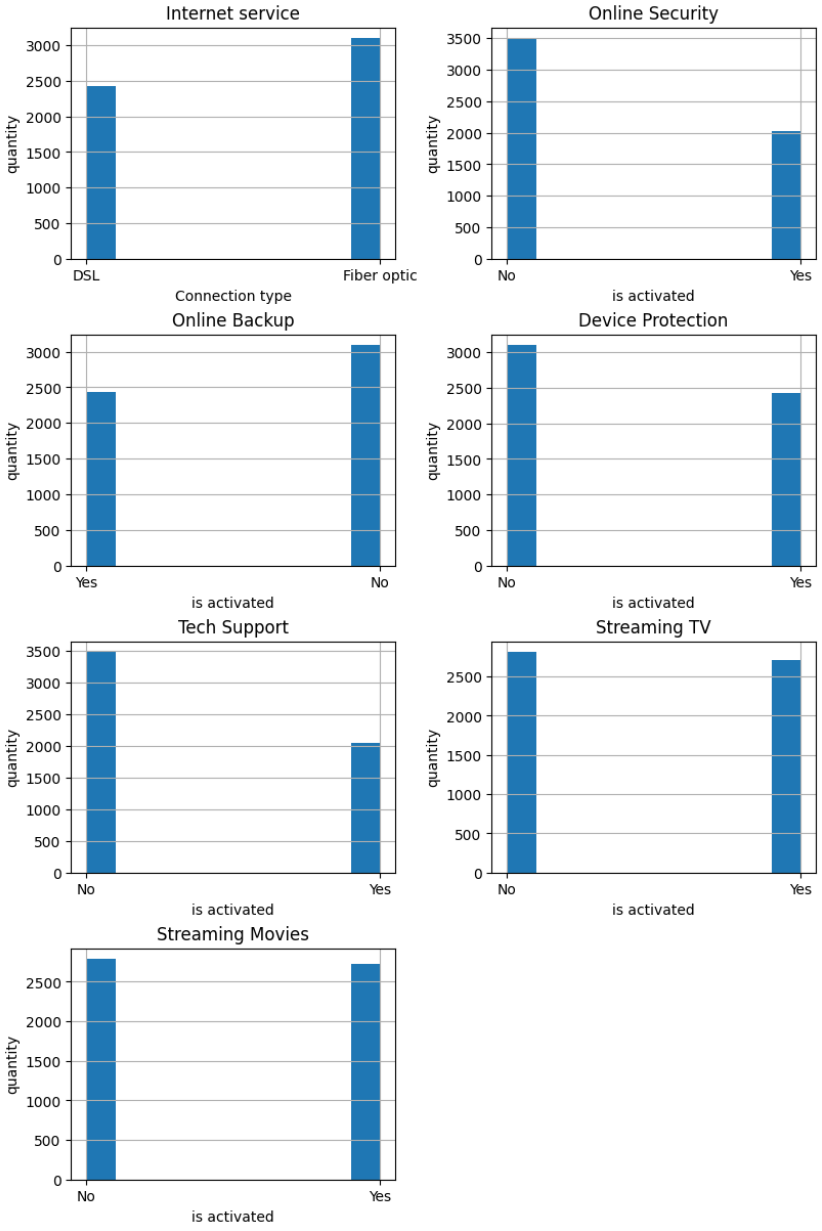

Рис. 8: распределения признаков интернет-услуг

3.4 Услуги телефонии

Таблица, представленная на рис. 9, состоит из 6361 строки, единственный столбец с категориальным признаком переключения на параллельную

customerID	MultipleLines
5575-GNVDE	No
3668-QPYBK	No
9237-HQITU	No
9305-CDSKC	Yes
1452-KIOVK	Yes

Рис. 9: Обзор таблицы с услугами телефонии

линию во время звонка.

Тип данных - строковый, пропусков в датасете нет. Распределение признака представлено на рис. 10

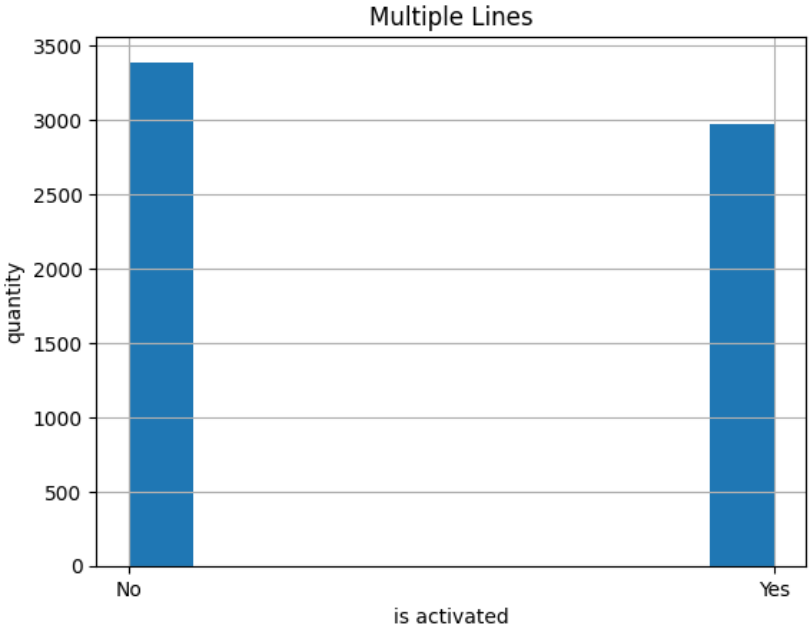

Рис. 10: Обзор таблицы с услугами телефонии

Пользователей без подключенной возможности переключаться между линиями больше на семь процентов

3.5 Выводы по разделу

Проведено чтение и первичный обзор данных в датасетах. В нашем распоряжении база из семи тысяч клиентов, пользующихся телефонией и/или интернетом. Явных пропусков нет.

Большинство признаков - качественные кроме двух: ежемесячного платежа и общей потраченной суммы. Что касается качественных признаков, данные в выборке распределены примерно равномерно. Что касается количественных признаков, распределение потраченной суммы стремится к Пуассоновскому, а у ежемесячного платежа два пика - в районе 20 (вероятно, соответсвующее минимальному ежемесячному платежу) и в районе 80-ти (возможно, это ежемесячный платёж популярного расширенного тарифа).

Распределение целевого признака изучено косвенно из даты окончания контракта. В выборке присутсвует дисбаланс: всего 15 процентов клиентов базы закончили сотрудничество с провайдером.

Характерный портрет пользователя в датасете - не имеющий пенсионного статуса, вероятнее всего - с иждевенцами, платящий ежемесячно 20 долларов в эконом-сегменте или 80-90 - в премиум-сегменте. Если говорить про интернет-услуги - с большей вероятностью это кабельное подключение.

Интернет-услуги в целом менее популярны телефонии: клиенты, пользующиеся услугой интернет-подключения, составляют 43 процента данных.

4 Формирование набора признаков

Данные по клиентам разнесены по нескольким таблицам, что неудобно для дальнейшего обучения моделей. Цель раздела - формирование единого набора признаков для дальнейшей аналитической работы.

4.1 Объединение таблиц и заполнение пропусков

Для дальнейшего обучения моделей на оплном наборе признаков произведено объединение таблиц. Слияние выполняется по уникальному полю идентификатора пользователя таким образом, чтобы включать в себя информацию о каждом клиенте. Выполнена проверка, представленная на рис. 11: ожидалось увидеть датасет из семи тысяч строк с признаками

```
Размер датасета = (7043, 20)
Количество значений и тип данных в столбцах:
<class 'pandas.core.frame.DataFrame'>
Index: 7043 entries, 7590-VHVEG to 3186-AJIEK
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   BeginDate         7043 non-null   datetime64[ns]
 1   EndDate           7043 non-null   object
 2   Type              7043 non-null   object
 3   PaperlessBilling  7043 non-null   object
 4   PaymentMethod     7043 non-null   object
 5   MonthlyCharges    7043 non-null   float64
 6   TotalCharges      7043 non-null   float64
 7   BeginYear         7043 non-null   int64
 8   InternetService   5517 non-null   object
 9   OnlineSecurity    5517 non-null   object
 10  OnlineBackup      5517 non-null   object
 11  DeviceProtection  5517 non-null   object
 12  TechSupport       5517 non-null   object
 13  StreamingTV       5517 non-null   object
 14  StreamingMovies   5517 non-null   object
 15  MultipleLines     6361 non-null   object
 16  gender            7043 non-null   object
 17  SeniorCitizen     7043 non-null   int64
 18  Partner           7043 non-null   object
 19  Dependents        7043 non-null   object
dtypes: datetime64[ns](1), float64(2), int64(2), object(15)
memory usage: 1.4+ MB
```

Рис. 11: Обзор результатов объединения таблиц

customerID	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	MultipleLines	gender
7469-LKBCI	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No	Male
8191-XWSZG	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No	Female
1680-VDCWW	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No	Male
1066-JKSGK	NaN	NaN	NaN	NaN	NaN	NaN	NaN	No	Male

Рис. 12: Пропуски в категориальных данных после объелинения таблиц

из всех таблиц, то есть должно получится 20 столбцов - с учётом уже созданного признака с годом. Результат проверки совпадает с предполагаемым.

В некоторых столбцах есть пропуски, являющиеся артефактом объединения данных (рис. 12): они связаны с тем, что не все не все клиенты пользуются и услугами телефонии, и интернетом. Пропуски как раз в столбцах, относящихся к этим секторам. Решено поступить следующим образом: если в этих столбцах пропуск - значит, услуга не подключена.

Для того, чтобы в дальнейшем можно было легко воспользоваться кодировщиком - заполним пропуски значением "No соответствующему данному факту в уже используемом наборе значений категориальных

признаков(рис. 13).

OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
No	Yes	No	No	No	No
Yes	No	Yes	No	No	No
Yes	Yes	No	No	No	No

Рис. 13: Категориальные признаки после заполнения пропусков

4.2 Добавление синтетических признаков

Первый признак - год подписания договора - уже добавлен.

Несмотря на то, что существенно дисбаланса категориальных признаков дополнительных услуг нет - есть гипотеза, что целевой признак может зависеть от агрегированного признака суммарного количества подключенных услуг. Ввиду этого в таблицу добавлен новый признак num_of_activated (рис. 14). Изучение его распределения проведено в разделе исследовательского анализа данных.

customerID	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	MultipleLines	num_of_activated
7590-VHVEG	No	Yes	No	No	No	No	No	2
5575-GNVDE	Yes	No	Yes	No	No	No	No	2
3668-QPYBK	Yes	Yes	No	No	No	No	No	2
7795-CFOCW	Yes	No	Yes	Yes	No	No	No	3
9237-HQITU	No	No	No	No	No	No	No	0
9305-CDSKC	No	No	Yes	No	Yes	Yes	Yes	4

Рис. 14: Демонстрация агрегированного признака количества подключенных услуг

Введён также ещё один признак - длительность контракта, вычисленная как разность даты окончания и начала договора (рис. 15). Для активных пользователей датой окончания является дата среза.

Явно выраженный целевой признак сформирован следующим образом: если даты окончания контракта нет - значит, клиент активный, значение целевого признака not_active равно 0, в противном случае - еди-

customerID	BeginDate	EndDate	contract_duration
7590-VHVEG	2020-01-01	2020-02-01	31
5575-GNVDE	2017-04-01	2020-02-01	1036
3668-QPYBK	2019-10-01	2020-02-01	123
7795-CFOCW	2016-05-01	2020-02-01	1371
9237-HQITU	2019-09-01	2020-02-01	153

Рис. 15: Демонстрация признака длительности контракта

нице. Затем ненужные столбцы - с датой начала и окончания договора - удалены.

В результате получена таблица с 21 столбцами из 7043 строк без пропусков, 7 признаков - количественные, остальные - категориальные.

4.3 Выводы по разделу

Таблицы признаков объединены в единый датасет признаков. В него вошли персональные данные по всем клиентам, параметры заключённых им договоров, а также сведения по всем подключенным услугам. Пропуски, являющиеся следствием объединения таблиц, заполнены.

Добавлены два признака: количество сервисов, подключенных у клиента, а также длительность контракта. Сформирован целевой признак, полученный путём преобразования столбца с датой окончания контракта: если даты нет - значит, клиент активный. Также сохранён признак с годом подписания контракта, добавленный на предыдущем шаге. Добавлен столбец с длительнностью контракта, сформированный как разность дат окончания и начала соотвественно. Для активных клиентов датой окончания является дата среза.

Более ненужные столбцы: с датами начала и окончания контракта - удалены.

5 Исследовательский анализ данных

Раздел посвящён детальному изучению признаков, поисков зависимостей и корреляций, а также прогнозированию важности признаков через призму прогнозирования оттока клиентов.

5.1 Изучение распределений признаков в разрезе целевого признака

Изучение распределений признаков для ушедших и активных клиентов Построены гистограммы распределений всех признаков в разрезе значений целевого признака (рис. 16).

Выделены следующие особенности:

- тип договора мало влияет на факт ухода из компании: несмотря на то, что в целом клиенты больше предпочитают помесячную оплату, среди ушедших клиентов виды договоров примерно равны. Среди клиентов, предпочитающих помесячный тип оплаты, процент ушедших ниже
- факт выставления счёта на электронную почту: среди ушедших клиентов больше тех, кто предпочитает такой способ - однако это соотвествует общему тренду в данных
- что касается методов оплаты, наиболее лояльны клиенты, придерживающиеся оплаты методом чека по почте. Возможно, это пожилые люди или просто консерваторы, предпочитающиеся стабильность, и потому хранящие верность одному провайдеру
- можно обратить внимание на разность распределений месячных расходов для уходящих клиентов: в целом основная масса клиентов сосредоточена в пике в районе 20 долларов (вероятно, ежемесячного платежа минимального тарифа) - и таких клиентов уходит довольно немного по сравнению с остающимися в этом ценовом сегменте. Примерно столько же по количеству уходит среди клиентов премиум-сегмента - тех, кто тратит более 80ти долларов в месяц - однако, это всего в три раза меньше оставшихся клиентов! То есть пользователи премиум сегмента - вероятно, те, кто пользуются расширенными пакетами и подключает дополнительные услуги, вероятно недовольны качеством сервиса.

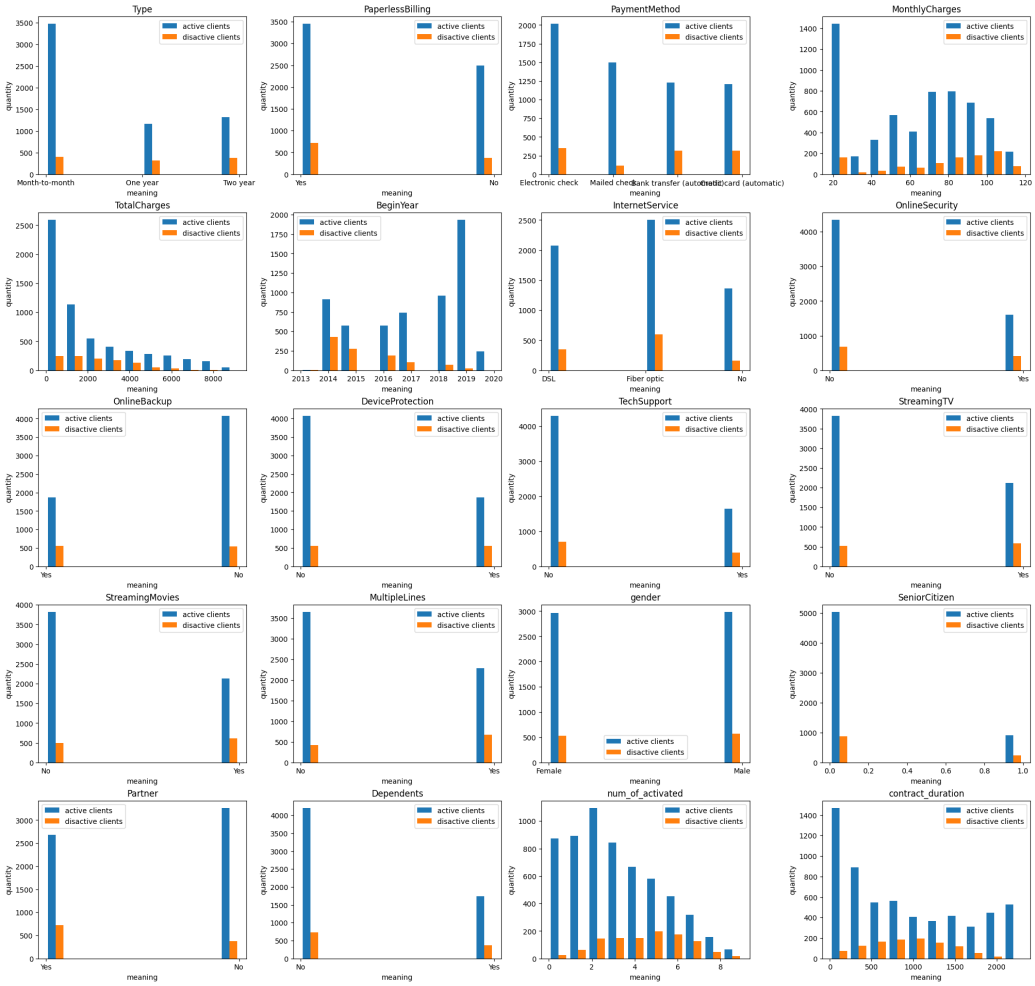

Рис. 16: Распределения характеристик клиентов в разрезе значения целевого признака

- что касается общих расходов, здесь наблюдается явная закономерность: чем больше клиент потратил, тем меньше шанс, что он уйдёт любопытна также засисимость оттока клиентов от года подписания договора: старые клиенты уходят существенно больше новых! Количество ушедших падает в зависимости от года. Возможно, старые клиенты пользуются устаревшими невыгодными для них тарифами, и потому уходят?

- клиенты, пользующиеся опттическим типом подключения, уходят чаще DSL - стоит обратить внимание на качество предоставления данной услуги
- среди подключенных услуг: антивируса, облачного хранилища, фильмов и тд, какого-то явного дисбалаанса нет
- мужчины уходят чуть чаще женщин: вероятно, ввиду большей склонности к оптимизации процессов
- люди с наличием партнёров более склонны уходить
- наличие иждивенцов мотивирует к постоянству - люди менее склонны уходить
- можно ожидать высокую важность признака с количеством подключенных услуг - распределения очень отличаются. Среди активных коиентов пик приходится на 2-3 подключенные услуги, а среди ушедших - на 5-6 подключенных услуг
- важность признака длительности вероятно тоже будет довольно высокой: видно, что активные клиенты в целом разделяются на две категории: пришедшие недавно или наоборот, на заре создания компании. Для ушедших клиентов распределение этого признака отличается качественно: в основном уходят клиенты, пользовавшиеся услугами данного оператора около 3ёх лет.

  В целом можно выделить следующую закономерность: основная масса активных клиентов - люди, которые пользуются минимальным тарифом с помесячной оплатой и небольшим количеством подключенных дополнительных услуг или вовсе без них. Большая часть клиентов, прекращающих пользоваться услугами компании - вероятно, клиенты премиум-сегмента с довольно большими ежемесячными расходами и большим количеством подключенных дополнительных услуг. Стоит отметить, что их договора в основном заключены довольно давно. Стоит обратить внимание на "старых"клиентов - может, пересмотреть их тариф и предложить более выгодный? Также возможно имеет смысл разработать пакеты более выгодного комплексного подключения услуг - в каждой конкретной услуге

нет явного дисбаланса в целевом признаке, однако уходящие клиенты в основном пользуются большим количеством допуслуг. Характерный срок оттока клиентов - три года пользования услугами компании.

В целом можно выделить следующую закономерность: основная масса активных клиентов - люди, которые пользуются минимальным тарифом с помесячной оплатой и небольшим количеством подключенных дополнительных услуг или вовсе без них. Большая часть клиентов, прекращающих пользоваться услугами компании - вероятно, клиенты премиум-сегмента с довольно большими ежемесячными расходами и большим количеством подключенных дополнительных услуг. Стоит отметить, что их договора в основном заключены довольно давно. Стоит обратить внимание на "старых"клиентов - может, пересмотреть их тариф и предложить более выгодный? Также возможно имеет смысл разработать пакеты более выгодного комплексного подключения услуг - в каждой конкретной услуге нет явного дисбаланса в целевом признаке, однако уходящие клиенты в основном пользуются большим количеством допуслуг. Характерный срок оттока клиентов - три года пользования услугами компании.

5.2 Зависимости признаков

Изучены корреляции между признаками (рис. 17). Большинтсво признаков - качественные, поэтому классическая корреляция линейной зависимости между признаками здесь не подойдёт. Использована библиотека phik.

Коэффициент корреляции для количественных признаков вычисляется как ковариация:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})((y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{1}$$

где N - размер столбцов, x_i, y_i - значения переменных для i-ого наблюдения.

Корреляция между категориальными признаками рассчитывается как:

$$\phi_C = \sqrt{\frac{\chi^2}{N \times min(r-1, k-1)}}, \tag{2}$$

где N - количество наблюдений, r и k - количество строк и столбцов соответственно. χ - результат статистического теста Пирсона между бинарными признаками, вычисляемый как:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{3}$$

где O_{ij} и E_{ij} - наблюдаемая и ожидаемая соответственно частоты ячейки для строки i и столбца j таблицы.

Обнаружены следующие зависимости:

- ежемесячные расходы существенно зависят от типа подключения интернета,подключения фильмов и стримингового ТВ. Также наблюдается корреляция с количеством подключенных услуг и суммарными расходами
- общая потраченная сумма коррелирует с количеством подключенных услуг и ежемесячгыми расходами, что логично
- те, кто подключают стриминговое телевидение, в основном подключают и каталог фильмов: видимо, это киноманы. Возможно, имеет смысл разработки выгодного пакетного предложения для этих услуг
- есть некоторая корреляция между наличием партнёра и наличием иждивенцев
- количество подключенных услуг существенно коррелирует не только с качественными колонками о факте подключения конкретной услуги и ежемесячными расходами, но и с наличием партнёра
- можно ожидать важность признака года подписания договора при дальнейшем анализе обученной модели
- признак длительности контракта сильно коррелирует с целевым признаком
- важно отметить, однако, высокую корреляцию признака длительности контакта с остальными признаками: характерна очень высокая корреляция с годом начала (близкая к единице!), высока корреляция с признаком общих расходов, присутствует некоторая корреляция с признаком типа подключения интернета.

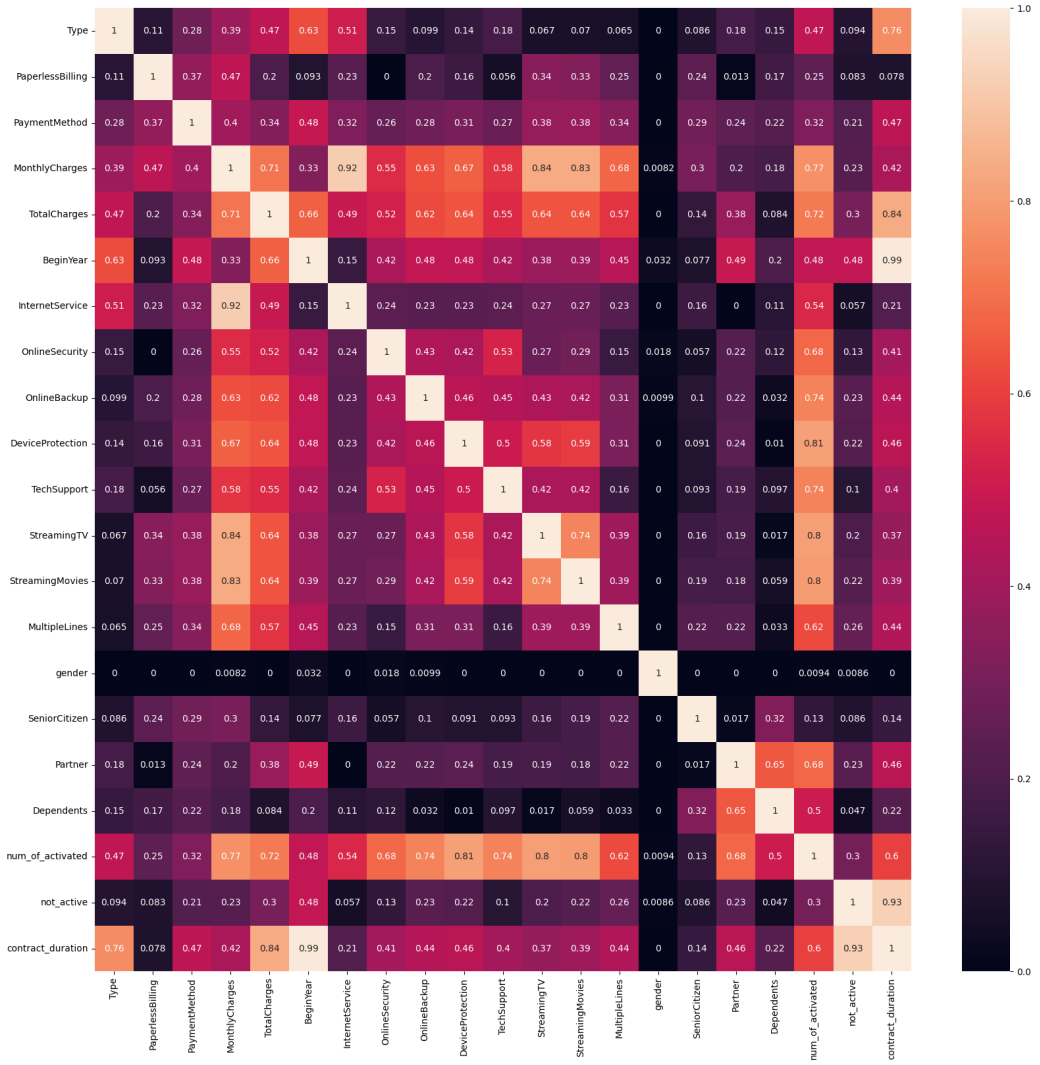

Рис. 17: Матрица корреляций признаков

5.3 Распределение признаков через призму года подписания договора

Дополнительно проведено изучение признаков по срезу значения года подписания договора: замечен пик количества договоров в 2019-ом году.

Изучены распределения ежемесяных расходов(рис. 18) и количества подключенных услуг(рис. 19) до 2019-ого года и начиная с 2019-ого.

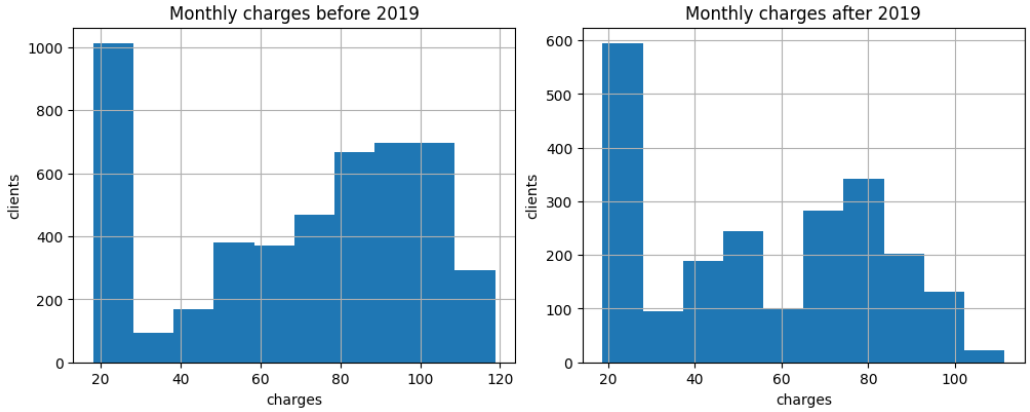

Рис. 18: Сравнение ежемесячных расходов пользователей с учётом порогового значения года подписания договора

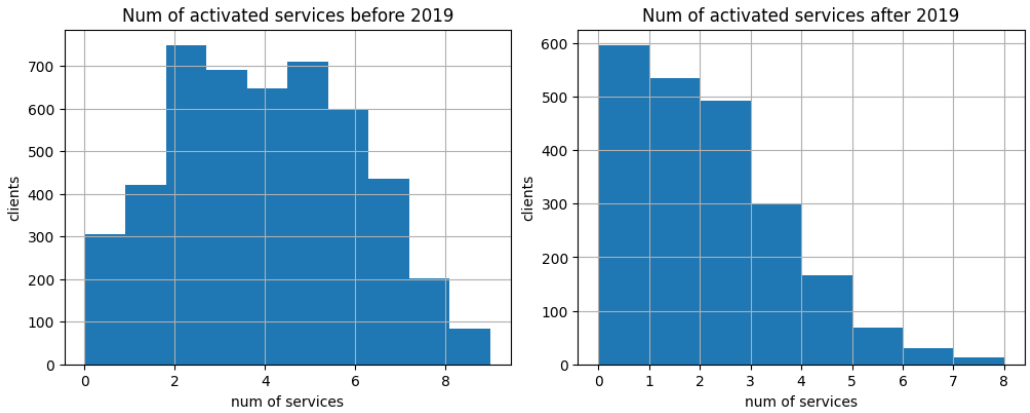

Рис. 19: Сравнение количества подключенных услуг с учётом порогового значения года подписания договора

Характер распределения по количеству сервисов выглядит совсем иначе, распределение ежемесяных платежей тоже отличается: то есть люди существенно чаще подключали дополнительные сервисы. Возможно, новые пользователи чаще пользуются новыми пакетами услуг, введёнными компанией в 2019ом году. Обнаружено, что ежемесячная прибыль от тех, кто пришёл в 2019ом году, больше почти в два раза (табл. 5.3).

Год подписания	Доход от новых клиентов, тыс $
2013	0.6
2014	68.7
2015	39.6
2016	37.6
2017	47.9
2018	58.6
2019	110.6
2020	9.2

значит, политика компании корректна и результативна.

5.4 Выбор набора признаков

В рамках выбора набора признаков для обучения моделей сравнивается год подписания договора и длительность контракта. Использование обоих столбцов некорректно, поскольку ведёт к утечке целевого признака. Корреляция целевого признака с длительностью выше, однако у длительности также очень высокая корреляция с признаками общих расходов и типа подключения интернета. Ввиду этого принято решение обучения выбранных моделей на двух вариантах наборов признаков (рис. 20) - с годом подписания договора и с длительностью контракта и выбором наиболее оптимальной конфигурации при сравнении полученных при обучении метрик.

```
data_duration = data.drop('BeginYear', axis=1)
data_year = data.drop('contract_duration', axis=1)
```

Рис. 20: Формирование обучающих наборов признаков

6 Подготовка данных для обучения моделей

6.1 Разделение данных на выборки

Сформированы два датасета, в одном из которых - признак года подписания договора, в другом - длительность контракта на момент среза.

Остальные признаки сохранены без изменений. Каждый из них разделён на обучающую и тестовую выборку в пропорциях 3:1 соответственно. Ввиду небольшого размера датасета принято решение не оформировать отдельную валидационную выборку и обучать модели с использованием кросс-валидации. Затем каждая из каждой полученной таблицы целевой признак сепарирован в отдельный датафрейм.

6.2 Масштабирование и кодирование признаков

Большинство признаков в датасете - категориальные, но со строковым типом данных в них. Для корректного процесса обучения моделей необходимо провести кодирование, т.е. преобразование в числовой тип. Будем работать с двумя кодировками - OneHotEncoder (рис. 21) для логистической регрессии и OrdinalEncoder (рис. 22) для моделей случайного леса и бустингов. OE выгоднее с точки зрения экономии вичислительных ресурсов за счёт сохранения количества признаков, но не подходит для линейных моделей - поэтому реализованы два рансформера для разных моделей.

```
ohe_transformer = make_column_transformer(
    (
        OneHotEncoder(drop='first', handle_unknown='ignore'),
        make_column_selector(dtype_exclude=np.number)
    ),
    (
        StandardScaler(),
        make_column_selector(dtype_include=np.number)
    ),
    remainder='passthrough'
)
```

Рис. 21: Реализация кодировщика OneHotEncoder с масштабированием количественных признаков

Присутствуют также количественные признаки: ежемесячные и общие расходы, год подписания договора, длительность контракта и количество подключенных услуг. Особенность заключается в существенном отличии диапазона значений признаков, поэтому в целях увеличения качества обучения проведено масштабирование данных, включенное в оба трансормера. Выбор столбцов реализован на базе ColumnTransformer с фильтром по типу данных в столбцах, что позволяет использовать кодировщик для всех вариантов наборов признаков. В OHE-кодировщике

```
oe_transformer = make_column_transformer(
    (
        OrdinalEncoder(),
        make_column_selector(dtype_exclude=np.number)
    ),
    (
        StandardScaler(),
        make_column_selector(dtype_include=np.number)
    ),
    remainder='passthrough'
)
```

Рис. 22: Реализация кодировщика OrdinalEncoder с масштабированием количественных признаков

первый столбец удаляется во избежание dummy-ловушки. В результате получены два обучающих набора преобразованных признаков - для OHE (рис. 23) и OE кодировщика (рис. 24).

	onehotencoder__Type_One year	onehotencoder__Type_Two year	onehotencoder__PaperlessBilling_Yes	onehotencoder__PaymentMethod_Credit card (automatic)
0	1.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0
2	0.0	0.0	1.0	0.0
3	1.0	0.0	1.0	0.0
4	0.0	0.0	1.0	0.0

Рис. 23: Пример данных с признаками, преобразованные с использованием OneHotEncoder

	ordinalencoder__Type	ordinalencoder__PaperlessBilling	ordinalencoder__PaymentMethod	ordinalencoder__InternetService
0	1.0	0.0	2.0	0.0
1	2.0	0.0	3.0	2.0
2	0.0	1.0	2.0	1.0
3	1.0	1.0	0.0	1.0
4	0.0	1.0	3.0	0.0

Рис. 24: Пример данных с признаками, преобразованные с использованием OrdinalEncoder

6.3 Борьба с дисбалансом

В распределении целевого признака (рис. 25) заметен существенный дисбаланс в сторону активных клиентов, что соответствует здравому смыслу и тенденциям сферы

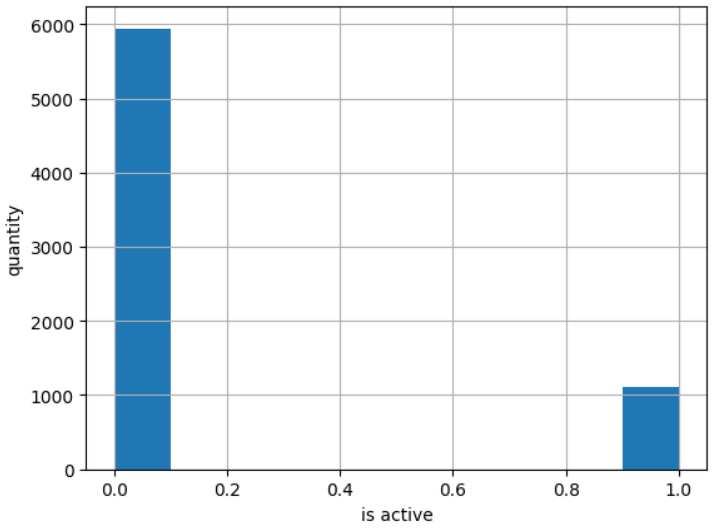

Рис. 25: Распределение значений целевого признака

То есть ушедшие клиенты составляют примерно 15 процентов датасета.

Из возможных методов борьбы с дисбалансом:

- уменьшение выборки: точно применять не будем, у нас и так довольно немного данных
- увеличение выборки: техника upsampling
- комбинация этих методов - SmotoTomek. В нашем случае выбранная метрика устойчива к дисбалансу, поэтому применять не будем взвешивание классов: будем применять в тех моделях, которые это позволяют сделать внутренними методами (например, в регрессии)
- изменение порога классификации.

Итого, сделаем так: применим взвешивание классов для тех моделей, где это возможно, а если метрика будет получаться недостаточно высокой - попробуем увеличение выборки и/или изменение порога классификации

6.4 Выводы по разделу

Сформированы два набора данных - с признаком года и признаком длительности, поскольку оставить оба признака одновременно нельзя. Данные разделены на обучающую и тестовую выборку в пропорции 3:1 соответственно. Ввиду небольшого количества данных отдельная валидационная выбока не выделяется - модели будут обучены с использованием кросс-валидации. Выделен целевой признак. Для признаков созданы трансформеры - OneHotEncoder для линейных моделей, OrdinalEncoder для "деревянных". В обоих случаях проведено масштабирование количественных признаков. Для борьбы с дисбалансом выбран метод взвешивания классов.

7 Построение, оптимизация и обучение моделей

Следующий шаг - построение моделей, способных предсказывать целевой признак. Обучение каждой модели проводилось на двух наборах обучающих данных в пайплайне с кросс-валидацией и оптимизацией гипераметров с целью поиска экстремума целевой метрики. Использование кросс-валидации позволяет также фитирование кодировщика на обучающей части выборки при каждой итерации.

Рассматриваются слеюущие модели:

- линейные модели: логистическая регрессия, модель Ridge, SGD классификатор;
- классификатор LightGBM;
- RandomForest;
- модели бустингов: градиентный бустинг и классификатор библиотеки CatBoost.

Для каждой пары датасет-модель получается значение целевой метрики на тренировочной выборке, время обучения, а также конфигурация гиперпараметров наиболее оптимальной модели.

Планируется сравнение пар датасет-модель для каждой модели, а после - выбор итоговой модели с учётом выбранного обучающего набора через призму значений метрики и времени обучения.

Идеологически обучение линейных моделей необходимо проводить на некоррелирующем наборе признаков, но пользуемся конструкторами моделей из бибилотиеки scikit-learn со встроенной регуляризацией, поэтому можно обучать на полном наборе признаков.

7.1 Логистическая регрессия

В основе классификатора логистической регрессии лежит сигмовидная функция, преобразующая линейную комбинацию входных признаков и из весов в вероятность принадлежности классу:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_1 - \beta_2 x_2 - ... - \beta_n x_n}}, \tag{4}$$

где Y - бинарный предсказанный результат, X - вектор признаков, $(\beta_0 ... \beta_n)$ - набор весов модели. Тогда P - вероятность принадлежности Y к классу 1 при зафиксированном значении X.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.77, время обучения =0.257;
- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.917, время обучения =0.198

7.2 Модель Ridge

Модель основана на снижении размерности пространства методом L2 регуляризации:

$$L_2 = \sum_i (y_i - y(t_i))^2 + \lambda \sum_i a_i^2, \tag{5}$$

- сумма квадратов весов модели, умноженных на гиперпараметр регуляризации. Здесь λ - настраиваемый гиперпараметр.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.77, время обучения =0.0404

- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.912, время обучения =0.0691

7.3 SGDClassifier

Классификатор стохастического градиентного спуска, основанный на методе опорных векторов и логистическо1 регрессии.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.764, время обучения =0.0799
- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.917, время обучения =0.0523

7.4 LightGBM

Метод градиентного бустинга, базирующийся на деревьях решений, при котором происходит расщепление только одного листа и использующий алгоритм, основанный на заполнении гистограмм.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.899, время обучения =0.661
- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.941, время обучения =3.36

Реализация классификатора средствами бибилиотеки scikit-learn позволяет ручное задание категориальных столбцов. Метрики для таких параметров конструктора модели:

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.899, время обучения =0.646
- для обучающего набора с годом подписания контракта: ROC-AUC =0.941, время обучения =0.5

Видно, что с определением категориальных признаков модель хорошо справляется самостоятельно - с точностью до тысячных разницы в метрике нет.

7.5 RandomForest

Алгоритм, являющийся ансамблем деревьер принятия решений. Прогноз является результатом агрегирования ответов множества деревьев.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.831, время обучения =1.05
- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.9, время обучения =1.04

7.6 Градиентный бустинг

Алгоритм, базирующийся на последовательной минимизации функции ошибки, стоящий предсказания в виду ансамбля более слабых предсказывающих моделей.

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.903, время обучения =5.25
- для обучающего набора с годом подписания контракта: ROC-AUC =0.944, время обучения =2.58

7.7 CatBoost

Классификатор библиотеки градиентного бустинга, имеющий внутренний кодировщик для категориальных переменных. Работа модели протестирована в двух режимах - с использованием встроенного и пользовательского кодировщиков. Результаты для этой модели с написанным кодировщиком:

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.901, время обучения =2.23
- для обучающего набора с годом подписания контракта: ROC-AUC =0.947, время обучения =1.71

При обучении модели на сырых данных с ручным заданием категориальных признаков:

- для обучающего набора с длительностью контракта: метрика ROC-AUC =0.866, время обучения =9.22

- для обучающего набора с годом подписания контракта: метрика ROC-AUC =0.945, время обучения =10.7

7.8 Сравнение моделей

Выбор наиболее оптимальной модели проведён в два этапа:

1. Сравнение целевой метрики для каждой рассмотренной модели, обученной на двух вариантах обучающий выборок - с годом подписания и с длительностью контракта (рис. 26)

2. Для выбранного обучающего набора рассмотрены две характеристики: метрика ROC-AUC и время обучения модели (рис. 27)

	Model	ROC-AUC duration	ROC-AUC year
0	LogRegression	0.770327	0.917383
1	Ridge	0.769509	0.911899
2	SGDClassifier	0.764323	0.916577
3	LightGBM	0.898886	0.940897
4	LightGBT_inner_transformer	0.898886	0.941375
5	RandomForest	0.830592	0.900198
6	GradBoosting	0.902757	0.944439
7	CatBoost_oe	0.874721	0.942227
8	CatBoost_kit	0.865966	0.945337
9	CatBoost_oe_optimized	0.900740	0.947020

Рис. 26: Целевая метрика обученных моделей в зависимости от набора признаков обучающей выборки

Хорошо видно, что на использовании обучающей выборки с признаком года подписания договора величина метрики получается выше, сравнимое качество обучения получается только для градиентного бустинга и модели CatBoost с оптимизацией гиперпараметров.

Важно отметить, что на выбор обучающего набора признаков существенное влияние оказывает предполагаемый временной горизонт жизненного цикла модели: в случае, если предполагается использовать мо-

	Model	ROC-AUC	fitting_time
0	LogRegression	0.917383	0.198050
1	Ridge	0.911899	0.069093
2	SGDClassifier	0.916577	0.052296
3	LightGBM	0.940897	3.355891
4	LightGBT_inner_transformer	0.941375	0.500302
5	RandomForest	0.900198	1.035187
6	GradBoosting	0.944439	2.578281
7	CatBoost_oe	0.942227	3.094604
8	CatBoost_kit	0.945337	10.676508
9	CatBoost_oe_optimized	0.947020	1.711599

Рис. 27: Целевая метрика и время обучения моделей на датасете с годом подписания контрактов

дель без переобучения и модификаций неограниченное время - длительность контракта надёжнее, если же акцент предполагается сделать на более точное прогнозирование в ближайший год-полтора - использование года даст более качественный результат.

В рамках текущей постановки задачи о максимизации целевой метрики остановимся на втором варианте.

Проведём сравнение всех обученных моделей и затраченного на обучение времени:

Наиболее важный критерий при выборе финальной модели - качество обучения, то есть значение полученной метрики, хорошо также учитывать время обучения, поскольку скорость работы модели для бизнес-задач часто важна, а вычислительные мощности могут быть ограничены.

Наиболее оптимальной является модель CatBoost с оптимизацией гиперпараметров и обучением на данных, преобразованных пользовательским кодировщиком - она показывает наиболее высокое качество обучения при конкурентноспособном времени обучения.

Соизмеримое качество метрики показывают модели градиентного бустинга - однако, они работает медленнее.

Дополнительное преимущество модели CatBoost - она показывает соизмеримо высокое качество на наборе признаков с длительностью контракта (ROC-AUC = 0.9 на обучающей выборке)

7.9 Выводы по разделу

Проведено обучение выбранных моделей на обучающей выборке с использованием кросс-валидации. Для каждой модели, если это возможно, проведена оптимизация гиперпараметров через призму максимизации выбранной метрики ROC-AUC. Получена комбинация гиперпараметров наиболее удачной конфигурации. Для каждой модели учитывалось также время выполнения. Наиболее оптимальной является модель CatBoost с подбором гиперпараметров, обученная на датасете с признаком года подписания договора, преобразованном пользовательским кодировщиком. В такой конфигурации ROC-AUC на обучающей выборке = 0.947.

8 Тестирование модели

8.1 Сравнение со случайной моделью

Тестирование полученной модели начнём со сравнения со случайной: если бы значение целевого признака выставилось случайно - 0 или 1.

Метрика ROC-AUC для случайной модели на обучающей выборке =0.509. Видно, что она существенно ниже полученной метрики для модели CatBoost - значит, применение модели рационально, поскольку она работает намного точнее случайного угадывания.

8.2 Изучение полученной модели

Теперь изучим полученную модель: возможно, её можно оптимизировать ещё. С этой целью изучим важность признаков - возможно, можно сократить количество признаков без ущерба для качества предсказаний и таким образом снизить нагрузку на вычислительные ресурсы. На финальной версии полученной модели для повышения интерпретируемости результата планируется нарисовать ROC-кривую, вычисление итогового

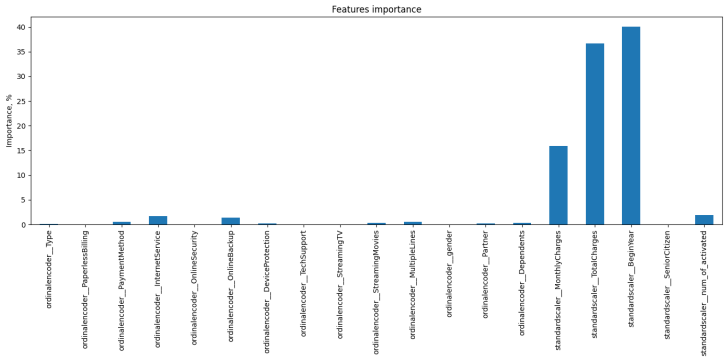

Рис. 28: Важность признаков финальной обученной модели

значения метрики на тестовой выборке, а также вычисление дополнительных метрик: точности, полноты и метрики f1.

8.2.1 Важность признаков

Изучена важность признаков предобученной модели (рис. 28) - возможно, часть признаков можно удалить, оптимизировав тем самым процесс обучения модели - увеличив метрику или уменьшив время обучения.

Выделены следующие особенности:

- общие расходы важнее ежемесячных в полтора раза, но оба признака очень важны. Корреляция между ними есть, но всё же не очень большая - порядка 0.7. Это значит, что можно оставить оба признака
- год подписания договора очень важен! Второй по важности признак после общих расходов
- признак количества дополнительных подключенных услуг существенно важнее категориальных столбцов с фактом подключения - значит, их можно попробовать удалить
- категориальный признак типа подключения интернета тоже важен
- признак года подключения ожидаемо важен
- факт выставления чека на электронную почту не важен, можно удалять

- важность типа подключения и метода оплаты соизмеримо не очень большая
- важность пола минимальна, что подверждает гипотезу из исследовательского анализа данных - в целом заивисимости целевого признака от пола. Признак можно удалить.
- Аналогичная важность - у признака наличия партнёра
- наличие иждивенцев влияет на целевой признак, хоть и совсем немного
- признак наличия пенсионного статуса не важен (на уровне десятых долей процента) - можно удалять этот признак
- признак количества подключенных услуг критично важен!

Итого: пробуем построить модель CatBoostRegressor на датасете со следующими признаками: общие и ежемесячные расходы, год подписания контракта, количество подключенных услуг, тип оплаты и подключения интернета, тип оплаты и наличие иждивенцев.

8.2.2 Построение модели CatBoost на сокращённом наборе признаков

Сокращено количество колонок в два раза, что позволило высвободить больше половины используемой памяти, занимаемой обучающим датасетом.

Засчёт того, что в трансформере выбор колонок реализован с использованием ColumnSelector по типу данных - новый трансформер писать не надо - а значит, можно пользоваться тем же пайплайном.

Запущен пайплайн с CatBoost классификатором - снова с подбором гиперпараметров, поскольку изменилось количество признаков. Кодировщик используется пользовательский - уже продемонстрировано, что в данном датафрейме он работает быстрее при сохранении качества предсказания.

Для модели CatBoost, построенной на сокращённом наборе признаков, метрика ROC-AUC =0.947, время обучения =2.0.

В целом сокращение набора признаков может быть целесообразно только с точки зрения экономии памяти, занимаемой датасетом - метрика не изменилась с точностью до тысячных, а время обучения возросло.

8.2.3 Вычисление финальных метрик на тестовой выборке

Конфигурация модели определена окончательно - это CatBoost классификатор в связке с пользовательским преобразователем данных, обученный на наборе признаков с годом подписания договора без длительности. Теперь можно вычислить итоговое значение целевой метрики такой модели на тестовой выборке, а также посчитать дополнительные метрики и построить ROC-кривую.

Конфигурация гиперпараметров итоговой модели, подобранная в пайплайне с кросс-валидацией:

```
cb_model_final = CatBoostClassifier(depth=2,
l2_leaf_reg=7,
learning_rate=0.05,
auto_class_weights='Balanced')
```

Метрика ROC-AUC для классификатора на тестовой выборке =0.958

Метрика f1 для классификатора на тестовой выборке =0.791

Такое значение означает, что модель хорошо прогнозирует класс 1, то есть что клиент собирается уйти.

Построена ROC-кривая(рис. 29) для контроля качества обучения модели:

Видно, что изгиб кривой высок - значит, модель хорошо классифицирует клиентов по категориям.

Отдельно выведена матрицу ошибок с конвертацией количества предсказаний в процент от общего числа. Обратим внимание на количество ложноотрицательных прогнозов: то есть случаев, когда модель не спрогнозировала уход клиента. Смотрим на эту характеристику отдельно, так как предложить промокод лояльному клиенту (потенциально недополученная прибыль) - это не так критично, как не удержать клиента, планирующего прервать сотрудничество.

```
[[79.27314026  4.6564452 ]
[ 2.49858035 13.57183419]]
```

Ложноотрицательная ошибка будет совершена моделью с вероятностью всего в 2,5 процента! То есть потенциально уходящего клиента модель пропустит с вероятнотью менее трёх процентов. С вероятностью в 4.6 процента рекомендательная система, основанная на этой модели, предложит промокод лояльному клиенту, который и не собирался никуда уходить.

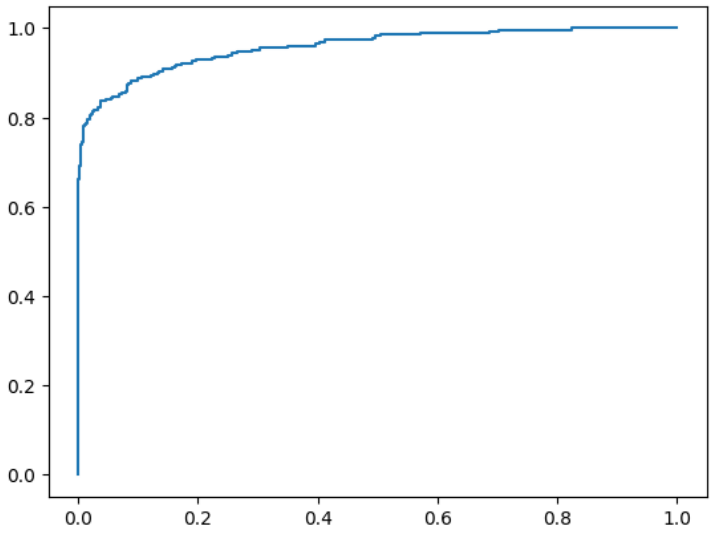

Рис. 29: Кривая обучения модели

9 Заключение

9.1 Обзор проекта и полученных результатов

Данный проект посвящён решению задачи классификации и прогнозированию оттока клиентов оператора связи "Ниединогоразрыва.com". Если модель определяет клиента как собирающегося уйти - планируется предложение клиенту скидочных промокодов и выгодных предложений. Входные данные представляют собой четыре датасета с информацией о клиенте и услугах, которыми он пользуется. Для оценки качества модели будет использована метрика ROC-AUC, в качестве дополнительной - метрика f1. Также при выборе модели будет учитываться время обучения.

Первая часть работы посвящена чтению и обзору данных. Выполнена загрузка данных и предворительный обзор каждого файла. Датасеты с информацией о клиентах и заключенных договорах содержат семь тысяч строк - по количеству клиентов, датасеты по услугам телефонии и интернета - 6 и 5,5 тысяч строк соответственно. Разница обусловлена количеством пользователей, пользующихся соответствующим сектором услуг. Явных пропусков в данных нет. Выполнен предварительный обзор каж-

дого датасета: изучены распределения, выбросы, особенности. Все признаки в этом датасете - качественные. Что касается особенностей личной информации, типичный абонент оператора - младше пенсионного возраста, имеющий иждивенцев, без явной выраженности в сторону какого-то пола и семейного положения. Преимущественно они предпочитают помесячный платёж, размер которого зависит от сегмента пользователя. Большинство пользователей составляют эконом-сегмент и платят около 20 долларов в месяц. Скорее всего, эта сумма соответствует ежемесячному платежу бюджетного тарифного плана. Однако, есть и клиенты премиум-сегмента - пик распределения их расходов находится в районе 80-100 долларов в месяц. Любопытным представляется распределение по году подписания договора - есть явный пик в районе 2019-ого года. Подробному изучению причинно-следственных связей посвящён следующий раздел. Проведено преобразование типов данных в столбце с общими расходами из строки в численный формат. Неявные пропуски заполнены нулевыми значениями, так как соответсвуют строкам с новыми клиентами, подписавшими договор в месяц выгрузки базы: данные об их первой платеже ещё отсутствовали в базе на момент выгрузки.

С целью формирования единого датасета признаков таблицы объединены таким образом, чтобы включить в себя данные обо всех предоставленных клиентах. Появившиеся пропуски образовались в категориальных столбцах с услугами интернета и/или телефонии. Пропуски заполнены таким образом, чтобы соответсвовать факту отсутствия подключения услуги.

Добавлены новые синтетические признаки: год подписания договора и суммарное количество подключенных клиентом услуг. Явным образом выявлен целевой признак из стобца с датой окончания действия договора: если даты нет - клиент активен и продолжает пользоваться услугами оператора, в противном случае клиент потерян. Затем ненужные столбцы: с датами начала и окончания контракта - удалены.

Следующий раздел посвящён исследовательскому анализу данных. Изучены распределения всех признаков в разрезе значения целевого признака. Обнаружено, что клиенты эконом-сегмента значительно более лояльны клиентов премиум-сегмента. Люди с наличием партнёра уходить чуть более склонны, а вот налиие иждивенцев больше мотивирует на стабильность. Можно ожидать высокую важность признака с количеством подключенных услуг - распределения очень отличаются. Среди активных коиентов пик приходится на 2-3 подключенные услуги, а сре-

ди ушедших - на 5-6 подключенных услуг. Также возможно имеет смысл разработать пакеты более выгодного комплексного подключения услуг - в каждой конкретной услуге нет явного дисбаланса в целевом признаке, однако уходящие клиенты в основном пользуются большим количеством допуслуг.

Проведена проверка на мультиколлинеарность. Видна взаимосвязь между ежемесячными и общими расходами, а также между расходами и типом подключения интернета. Те, кто подключают стриминговое телевидение, в основном подключают и каталог фильмов: видимо, это киноманы. Возможно, имеет смысл разработки выгодного пакетного предложения для этих услуг. Наличие партнёра связано не только с наличием иждивенцев, но и с суммарным количеством подключенных услуг. Дополнительно изучено распределение платежей и количества подключаемых услуг в разрезе года подписания договора. В качестве порогового значения выбран 2019ый год - год, когда заметен явный прирост клиентов. Видно, что в конце 2018ого-2019ого компания, вероятно, существенно поменяла стратегию и стала больше ориентироваться на привлечение бОльшего количества клиентов эконом-сегмента. Исходя из того, что клиенты с платежами в 80-100 долларов сохранились, а количество дополнительных услуг существенно сократилось - можно предположить, что компанией были пересмотрены тарифные планы и введено больше пакетных предложений. Ежемесячная прибыль от тех, кто пришёл в 2019ом году, больше почти в два раза - стратегия компании отлично сработала.

Следующий этап заключается в подготовке датасета для дальнейшего обучения моделей. Ранные разбиты на признаки и целевой признак, а затем - на обучающую и тестовую выборку в пропорциях 3:1 соответственно. Ввиду небольшого размера исходных данных валидационная выборка отдельно не выделяется, запланировано обучение моделей с использованием кросс-валидации. В датасете присутсвуют количественные признаки, котореы необходимо отмасштабировать, и качественные, которые нужно закодировать. Поскольку планируется обучение моделей различной конфигурации - линейные, случайный лес, бустинги - используется два варианта преобразования категориальных признаков - OneHotEncoder и OrdinalEncoder. В датасете присутствует существенный дисбаланс целевого признака: ушедшие клиенты составляют всего 15 процентов общего датасета. В качестве борьбы с дисбалансом выбран метод взвешивания классов внутренними методами конструктора моде-

ли.

Обучены следующие модели: логистическая регрессия, модель Ridge, SGD классификатор, LightGBM, модель случайного леса (отдельное дерево пробовать не будем), градиентный бустинг, CatBoost. Пайплайн включает в себя кросс-валидацию - это позволит в том числе обучать кодировщик на только на обучающей выборке на каждой итерации кросс-валидации. В рамках каждой модели - там, где это представляется возмодным - проведен подбор гиперпараметров с целью поиска экстремума целевой метрики. Классификатор CatBoost, показавший хорошее качество предсказания с заводскими настройками, обучен в двух конфигурациях - с использованием пользовательского кодировщика признаков и в варианте обучения модели на сырых данных с использованием внутренних кодировщиков модели. Продемонстрировано, что качество предсказания у таких моделей соизмеримо, однако скорость обучения модели на сырых данных ниже в три раза. На выходе для каждой модели получены значения метрики ROC-AUC на обучающей выборке и времени обучения лучшей конфигурации модели. Все значения продемонстрированы в сводной таблице. В качестве наиболее оптимальной выбрана модель CatBoostClassifier с использованием пользовательского кодировщика и подбором гиперпараметров - такая модель демонстрирует максимальное значение метрики при конкурентноспособном времени обучения.

Проведено тестирование полученной модели методом сравнения её с константной: предобученный классификатор предсказывает факт потенциального ухода клиента из компании существенно лучше наивного предиктора: целевая метрика классификатора практически в два раза выше.

Обучение каждой модели проведено в двух вариантах конфигураций - на датасете с признаком года подписания договора и с использованием в качестве признака длительность контракта. Использование обоих характеристик одновременно невозможно, поскольку ведёт к утечке целевого признака.

Изучена важность признаков получившейся модели: как и прогнозировалось, агрегированный количественный признак количества подключенных услуг важнее категориальных признаков по каждой услуге. Важны признаки расходов, введение признака года подписания договора оправдало себя: это второй по важности признак. Проведено дополнительное обучение классификатора CatBoost на сокращённом количестве признаков. Метрика изменяется в тысячных долях, а время обу-

чения немного увеличивается. Данный подход может быть использован при ограничении в памяти - сокращённый датасет занимает в три раза меньше места при сохранении качества предсказаний. Однако, в условиях текущего ТЗ выбрана стратегия обучения модели на полном наборе признаков.

Финальный этап - вычисление метрик обученной модели на тестовой выборке. Получившиеся результаты:

- ROC-AUC = 0.96
- f1-score = 0.89

ROC-кривая демонстрирует, что модель хорошо справляется с разделением клиентов по категориям построена матрица ошибок: клиента, планирующего расторгнуть договор, модель пропустит с вероятностью в 2.5

9.2 Анализ реализации поставленной ТЗ

Исходное ТЗ проекта - прогнозирование оттока клиентов с целью предложения ему промокодов, выгодных предложений и тд. Пороговое значение целевой метрики ROC-AUC = 0.85. Достигнутое значение метрики превосходит порог на 11%. Полученная конфигурация модели имеет высокое качество предсказаний по f1-метрике, что позволяет говорить о том, что модель качественно решает поставленную задачу.

9.3 Дальнейшее развитие проекта

Поскольку локальной моделью была максимизация целевой метрики - выбран набор признаков с годом подписания договора, однако в долгосрочной перспективе промышленного использования модель, обученная на сроке контракта, выглядит перспективнее ввиду удобства и большей универсальности. Поэтому можно интересным выглядит перспектива оптимизации такой модели. Также интересно было бы выстроить рекомендательную систему, основанную на ансамбле независимых предобученных моделей.

В рамках уже имеющейся модели возможно расширить проект, добавив более глубокий статистический анализ зависимости признаков для фильтрации набора признаков в тренировочный датасет. Также можно

было бы протестировать изменение качества предсказания в зависимости от методов борьбы с дисбалансом (например, при применении методики Smote Tomek).