Отчет

Выгрузка включала в себя 4 таблицы:

1. таблицу, содержащую информацию о контракте:
   1.1. дата заключения договора
   1.2. дата расторжения договора
   1.3. тип подключения (ежемесячное, ежегодное, подключение на 2 года)
   1.4. использование электронного счета (да, нет)
   1.5. способ оплаты (электронный чек, чек по почте, автоматический банковский перевод, автоматичекий перевод с кредитной карты)
   1.6. размер ежемесячного платежа
   1.7. размер суммарного платежа
2. таблицу, содержащую информацию об интернет-услугах:
   2.1. интернет сервис (ДСЛ, оптоволокно)
   2.2. онлайн безопасность (да, нет)
   2.3. резервное копирование (да, нет)
   2.4. защита устройства (да, нет)
   2.5. тех. поддержка (да, нет)
   2.6. стриминговое тв (да, нет)
   2.7. стриминговые фильмы (да, нет)
3. таблицу, содержащую персональные данные клиента:
   3.1. пол (м, ж)
   3.2. пенсионер (да, нет)
   3.3. партнер (да, нет)
   3.4. зависимые (да, нет)
4. таблицу, содержащую информацию о телефонных услугах:
   4.1. поключение нескольких линий (да, нет)

Общая таблица состояла из 7043 строк и содержала пропущенные значения (связано с тем, что данные услуги не были подключены):

1. интернет_сервис
2. онлайн_безопасность
3. резервное_копирование
4. защита_устройства
5. тех_поддержка
6. стриминг_тв
7. стриминг_фильмы
8. несколько_линий

не подключена 1 улсуга (682): несколько_линий
не подключено 7 услуг (1526): интернет_сервис, онлайн_безопасность, резервное_копирование, защита_устройства, тех_поддержка, стриминг_тв, стриминг_фильмы

все отсутствующие значения были заполнены пометкой "нет"

Целевая переменная:

Характеристика потока клиентов:

1. первичный наплыв клиентов: 2014-02-01
2. период с оттоком: 2019-10-01:2020-01-01
3. последний месяц с подключениями в оттоке не учитывается (новые клиенты): 2020-02-01
4. с 2019-09-01 наблюдается интесивное появление новых клиентов

Ушедшими(~30%) клиентами считались те, кто:

1. подключился хотя бы за месяц до последней даты выгрузки
2. расторгнувший договор

из-за наличия дисбаланса классов в целевой переменной, ценка моделей проводилась по стратифицированным разбиениям данных

Модели:

1. линейные (LinearSVC, LogisticRegression)
2. метрические (KNN)
3. деревянные (DecisionTree, ExtraTree, RandomForest, ExtraTrees, Lightgbm)

Пайплайны:

Так как использовались разные типы моделей, пайплайны подготовки данных тоже были разными:

1. для линейных моделей категории кодировались с помощью дамми-кодирования и с помощью целевой переменной; производилось масштабирование
2. для метрических моделей категории кодировались с помощью лейбла и с помощью целевой переменной; производилось масштабирование
3. для деревянных моделей категории кодировались с помощью лейбла и с помощью целевой переменной; производилось масштабирование

Были сформированы различные комбинации кодирований (для выявления переобучения на конкретный)

Метрика качества:

1. отбор признаков и оптимизация гиперпараметров производились для: accuracy, f1, precision, recall
2. модели, переобученные на тренировочные данные, исключались

модель	метрика	валдиационная метрика	отложенная метрика	ложные срабатывания	ложные пропуски
svm	recall	0.94	0.95	132	23
knn	precision	0.91	0.93	30	91
lgb	accuracy	0.86	0.89	53	24

Я бы боролся за удержание клиентов. Это положительно влияет на имидж компании, а так как Яндекс имеет множество монетизирующихся сервисов, удержание клиентов является более предпочтительным (svm+recall: 0.95 на тестовой выборке)

Оценка финальной модели:

Наиболее важными являются признаки:

1. длительность_подключения(обратная зависимость от вероятности оттока)
2. стриминг_фильмы(обратная зависимость от вероятности оттока)
3. электронный_счет(прямая зависимость от вероятности оттока)

при большем размере выборки стоит построить ансамбль моделей