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. деревянные: Lightgbm,RandomForest, ExtraTrees, DecisionTree, ExtraTree
2. линейные: LogisticRegression
3. метричекские: KNN

Пайплайн:

категории кодировались с помощью лейбла и с помощью целевой переменной; производилось масштабирование

Метрика качества:

- roc_auc

model	cv	te
Lightgbm	0.9633857030918538	0.9525232987515385
RandomForest	0.955295353895508	0.9525232987515386
ExtraTrees	0.9540459425803987	0.9547471529494512
DecisionTree	0.9504379612676358	0.9431107065650245
ExtraTree	0.9151612081300116	0.8586145905522398
LogisticRegression	0.9574972921507132	0.952926695559532
KNN	0.9296703529405397	0.9307398711198915

Финальная модель

- модель

ExtraTreesClassifier(max_depth=12, min_samples_leaf=2, n_estimators=1000, n_jobs=-1, random_state=13)

- метрики

metric	cv	te	fp	fn
auc	0.954046	0.954747	NaN	NaN
accuracy	0.912295	0.903409	437	118

Наиболее важными являются признаки:

1. длительность_подключения□_v2 (остаток от дедения длительность_подключения на 12, закодированный целевой переменной)
2. тип_подключения<->способ_оплаты (категория тип_подключения и способ_оплаты)
3. длительность_подключения□_v2 (остаток от дедения длительность_подключения на 12)
4. ежемесячный_платеж_log_round0_v2<->суммарный_платеж_log_round0_v2 (клмбинации округленных логарифмов суммарного и ежемесячного платежей)
5. суммарный_платеж_log_round2_v2 (округленный логарифм суммарного платежа)

Дальнейшие исследования:

При большем размере выборки стоит построить ансамбль моделей

Ответы на вопросы

1. Какие пункты плана были выполнены, а какие — нет (поясните почему)?

Все пункты плана были выполнены

2. Какие трудности возникли и как вы их преодолели?

В целом, трудностей не возникало

3. Какие ключевые шаги в решении задачи выделили?

Определить целевую переменную

4. Какая ваша итоговая модель и какое у неё качество?

- ExtraTreesClassifier(max_depth=12, min_samples_leaf=2, n_estimators=1000, `n_jobs=-1, random_state=13`)
- auc на тестовой части равен 0.954747 |metric|cv|te|fp|fn| |---|---|---|---|---| |auc|0.954046|0.954747|NaN|NaN| |accuracy|0.912295|0.903409|437|118|