**Описание решения:**

1. Загрузка данных.
2. Разбиение на тренировочную и тестовую выборки.
3. Детальное изучение данных:
   * Ввиду сильной корелляции между признаками POLICY\_BEGIN\_MONTH и POLICY\_END\_MONTH, можно сделать вывод, что полисы в подавляющем большинстве выдаются на целое число лет. В то же время, сезон выдачи полиса вряд ли как-то влияет на принятие клиентом решения о продлении полиса, так что данные признаки рассматривать не будем.
   * POLICY\_SALES\_CHANNEL, POLICY\_SALES\_CHANNEL\_GROUP - категориальные признаки. POLICY\_SALES\_CHANNEL принимает достаточно много различных значений, при этом POLICY\_SALES\_CHANNEL\_GROUP - это сгруппированные значения данного признака. В связи с этим POLICY\_SALES\_CHANNEL рассматривать не будем, а POLICY\_SALES\_CHANNEL\_GROUP необходимо будет разбить на несколько признаков по категориям, используя one hot encoding.
   * POLICY\_BRANCH - категориальный признак, принимающий 2 значения. Закодируем значения признака числами.
   * POLICY\_MIN\_AGE - числовой признак. Конкретный возраст вряд ли сильно кореллирует с вероятностью пролонгации полиса. Однако если разбить людей по возрастным группам (например, "18-25 лет", "25-40 лет", "40-65 лет", "65+ лет"), можно будет проследить зависимости.
   * POLICY\_MIN\_DRIVING\_EXPERIENCE - числовой признак. Вероятно, в некоторых местах указан год начала вождения вместо опыта. Исправим (будем считать, что данные за 2016 год, судя по времени создания файла с описанием задания). Кроме того, лучше выделить несколько групп водителей с разным стажем, аналогично тому, как это сделано для возраста.
   * VEHICLE\_MAKE и VEHICLE\_MODEL - категориальные признаки. Их кодирование с помощью one hot encoding даст нам большое количество признаков со множеством нулей. Однако решение о пролонгации полиса зависит от марки и модели машины, поэтому попробуем их оставить и применим к ним one hot encoding (результаты тестирования над данными с учетом данных признаков дают прирост оценки accuracy на 0.05, так что оставим их при обучении итоговой модели).
   * VEHICLE\_IN\_CREDIT – категориальный признак, значения 0 и 1, оставляем как есть.
   * VEHICLE\_ENGINE\_POWER, VEHICLE\_SUM\_INSURED – числовые признаки, необходимо отмасштабировать.
   * POLICY\_INTERMEDIARY - категориальный признак с 1333 различных значений. Обойдемся без него.
   * INSURER\_GENDER – категориальный признак, принимает 2 значения. Сделаем из него бинарный признак со значениями 0 и 1.
   * POLICY\_CLM\_N, POLICY\_CLM\_GLT\_N, POLICY\_PRV\_CLM\_N, POLICY\_PRV\_CLM\_GLT\_N – категориальные признаки, необходим one hot encoding В данных признаках неопределенные значения обозначаются по-разному (как N и n/a). Необходимо привести к общему значению.
   * CLIENT\_HAS\_DAGO, CLIENT\_HAS\_OSAGO, POLICY\_COURT\_SIGN, POLICY\_HAS\_COMPLAINTS - бинарные признаки, обработка не требуется.
   * CLAIM\_AVG\_ACC\_ST\_PRD – числовой признак, необходимо масштабирование.
   * POLICY\_YEARS\_RENEWED\_N - числовой. В 46 примерах не определено значение. Установим для данных примеров значение 1, как наиболее часто встречающийся в выборке вариант.
   * POLICY\_DEDUCT\_VALUE – числовой признак, потребуется масштабирование.
   * CLIENT\_REGISTRATION\_REGION – категориальный признак, 83 значения. Закодируем через оne hot encoding.
   * POLICY\_PRICE\_CHANGE – числовой признак, потребуется масштабирование.
4. Применение операций удаления, масштабирования и кодирования признаков для тренировочного набора данных.
5. Разбиение тренировочной выборки на новую тренировочную и валидационную (для оценки качества модели).
6. Обучение модели.

***Комментарий****:* Для обучения было опробовано несколько моделей(логистическая регрессия, случайный лес, градиентный бустинг, нейронная сеть с одним скрытым слоем) с различными параметрами. Качество модели оценивалось по метрике accuracy (т.к. в задании требуется максимизировать кол-во правильно классифицированных объектов, то есть TP+TN). Результаты показали, что лучше всего на имеющихся данных показывает себя градиентный бустинг. Параметры обучения подбирались по сетке.

1. Применение операций удаления, масштабирования и кодирования признаков для тестового набора данных.
2. Получение прогнозов и вероятностей модели для тестовой выборки.
3. Запись полученных результатов в csv-файл.

Полученная в итоге модель - лучшая из всех, с которыми были проведены эксперименты. Улучшить качество теоретически можно при построении модели по сетке с большим числом и разбросом гиперпараметров. Другой способ – изменение качественного и количественного состава участвующих в обучении признаков.