Рекомендации услуг Мегафон

Гладков Антон

Описание задачи

Телекоммуникационные компании сейчас существуют не только за счет базовых услуг по предоставлению связи, наподобие обычных звонков и СМС, а уже давно расширяют арсенал возможностей:

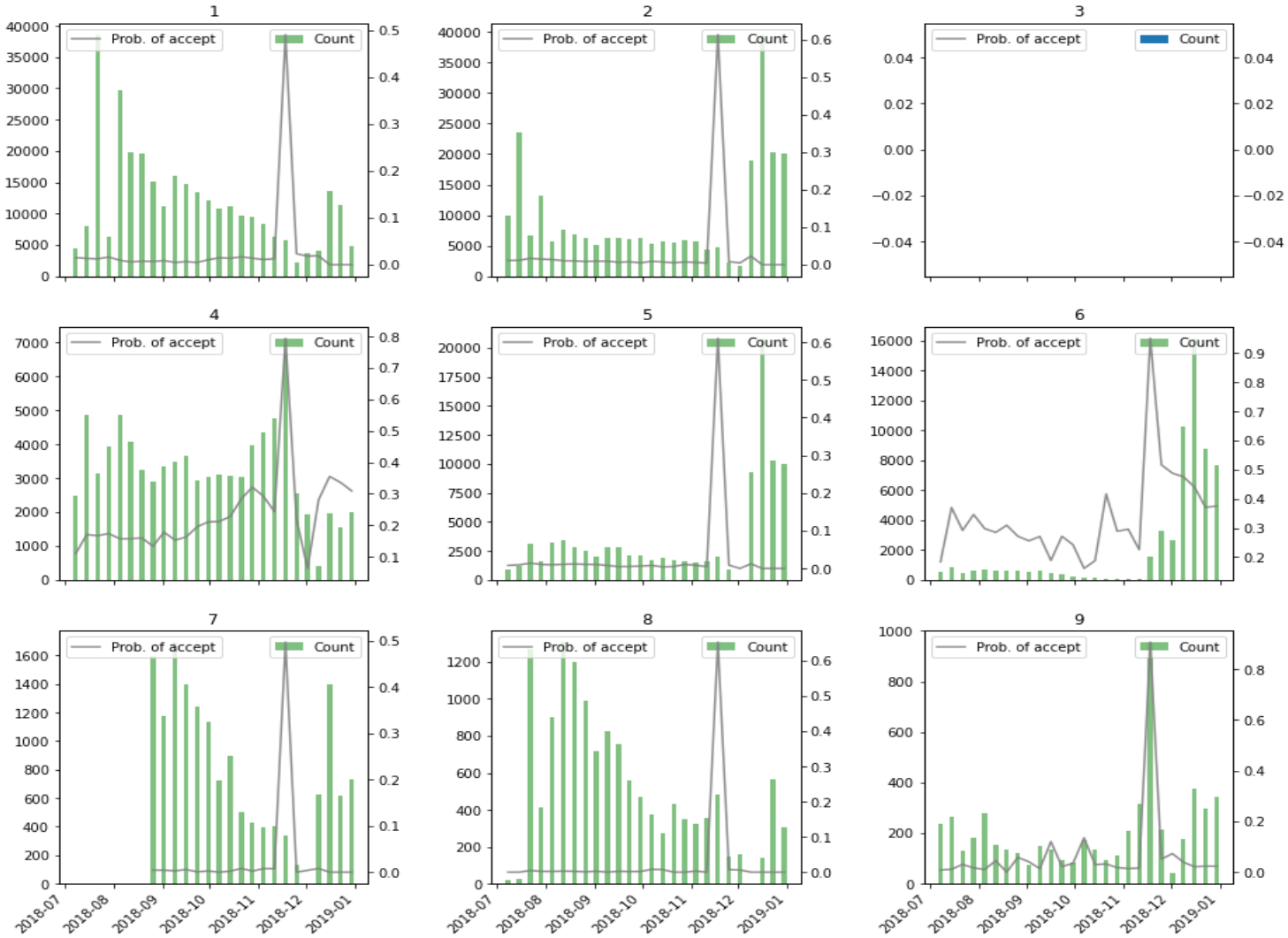
- во-первых, для создания клиентской базы, которая будет меньше нуждаться в сторонних решениях (больше пользоваться киносервисом 'START' вместо 'Кинопоиска' и т.д.)
- во-вторых, для поиска новых путей получения прибыли и сфер влияния.

Именно поэтому важно привлекать новых клиентов к наилучшим предложениям (а также старых уже проверенных абонентов к новым решениям), но делать это необходимо не безраборно, а с максимальной выгодой для компании. Для решения этой проблемы ниже будет рассмотрен алгоритм предложения различных услуг клиентам "Мегафона" в различные временные периоды.

Обзор данных

В наличии имеется 8 услуг, имеющие разные вероятности подключения после предложений клиентам.

- Четко можно выделить групп (2, 5, 6 услуги) с возросшим количеством предложений с определенной даты.
- Для всех услуг есть неделя, когда процент подключений был подозрительно высоким - с 2018-11-18 (эту неделю из обучения пока предлагается убирать).
- Только 4 и 6 имеют сильно отличные от нуля вероятности подключения.



Проблемы датасета

Проблема 1:

- Наличие "странной" недели с 2018-11-18 с большой вероятностью подключения

Решение:

- Убрать неделю из датасета. При возможности дополнительно анализировать его отдельно при наличии дополнительных данных оператора

Проблема 2:

- Из-за наличия только исторических данных с определенной даты для старых записей для обучения скорее всего не будет актуального статуса по подключенным услугам
- Нет данных по отключенным услугам

Решение:

- Необходим доступ к первоначальным данным на HLR/HSS для тех записей, для которых не сохранилось истории

Проблема 3:

- Очень крупная агрегация по времени (недельная), из-за чего невозможно анализировать поведение абонентов в разные часы/дни

Решение:

- Необходим сбор с более плотной агрегацией. В идеале данные нужно обновлять каждый час

Сравнение моделей

Для реализации задачи было построено 3 модели (1 на основе стат. подхода, еще 2 - ML):

Статистический подход	Лог. регрессия	**Catboost**
Метки - вероятности подключения услуг на основе данных для обучения. Чтобы внести вариативность, метки умножались на случайные числа из нормального распределения с 1 дисперсией.	Сравнивалась 1 модель бинарной классификации с совокупностью моделей для каждой из услуг. Первый вариант дал более хороший результат.	Модель дала самый лучший результат и превысила бейзлайн, основанный на базовой статистике. Предположительно, в данных содержатся достаточно сложные нелинейные зависимости, из-за чего лог. регрессия не выделилась
Метрика **F1 – 0.698**	Метрика **F1 – 0.549**	Метрика **F1 – 0.723**

Посткорректировка результатов

В виду того, что анализировать по времени в данном случае можно было крайне ограниченно, а данных по каждому абоненту было представлено крайне мало (менее 1% от всех пользователй обладают хотя бы двумя строками в выгрузке), было решено корректировку результатов делать только в части того, чтобы не рекомендовать услугу клиенту, если она ранее уже была предложена (не важно с каким резальтатом, потому как если он отказался, то ему она не интересна, а если принял, то она у него уже есть).

Что можно улучшить

- Более плотно поработать с выбором параметров для модели. Т.к. планируемый бизнес результат достигнут, дальнейшая оптимизация не производилась
- Для честной оценки лог. регрессии можно было использовать нормализацию данных, однако при предварительной оценке сильного разброса по весам не наблюдалось (за исключением даты, которая была исключена в данном случе)
- Применить рекомендации для решения проблем датасета и повторить эксперименты