Задача прогнозирования оттока клиентов.

Актуальность и цели задачи.

Для дипломной работы была выбрана **задача прогнозирования оттока клиентов** телекоммуникационной компании. Эта **задача заключается** в анализе данных клиентской базы и прогнозировании вероятности того, что клиент перестанет пользоваться сервисом в течение некоторого заданного заранее промежутка времени, например, решит перейти к конкуренту или откажется от использования услуги данного типа вообще.

Актуальность данной задачи обусловлена тем, что в России рынок телекоммуникационных услуг достаточно конкурентен и снижение оттока даже небольшого числа абонентов может значительно улучшить финансовые показатели компании и отразиться на ее доле на рынке. На картинке показаны доли крупнейших российских сотовых операторов в 2006-2016 гг.

В целом, задача предсказания оттока клиентов находит свое применение не только в области телекоммуникаций, но и в банковском секторе, страховых компаниях и других организациях, предоставляющих услуги в сегменте B2C.

Целью проекта является построение математической модели (бинарного классификатора), которая на основе имеющихся данных, предсказывает вероятность оттока пользователя (т.е. относит пользователя к классу «отток» или «не отток» с определенными вероятностями). **Целевым значением** модели является вероятность отнесения пользователя к классу «отток» (т.е. чем этот показатель выше, тем больше шансов, что данный пользователь уйдет). Важно заранее находить пользователей, склонных к оттоку, чтобы вовремя предотвратить отток и провести кампанию по удержанию клиентов (в частности, выявить и устранить причины оттока и т.д.), т.к. чаще всего расходы на привлечение новых клиентов превышают расходы на удержание уже имеющихся.

На практике данный классификатор поможет определить эффективную долю пользователей, склонных к оттоку, которых следует взять в программу по удержанию клиентов, для максимизации экономического эффекта от проведения данной компании.

Описание данных

Данные представляют собой данные французской телекоммуникационной компанией Orange. В задаче речь идет о клиентских данных, поэтому данные были предварительно обфусцированы и анонимизированны: из датасета убрана любая персональная информация, позволяющая идентифицировать пользователей, а также не представлены названия и описания переменных, предназначенных для построения прогнозов. Набор данных, приведенных в задаче, состоит из 50 тыс. объектов и включает 230 переменных, из которых первые 190 переменных - числовые, и оставшиеся 40 переменные - категориальные.

На первом этапе решения задачи был **проведен описательный анализ и визуализация данных.** В частности, с помощью коэффициента Крамера (т.к. целевая функция представляет собой бинарную переменную) была сделана предварительная оценка важности признаков, были выделены признаки, которые возможно окажут наибольшее влияние на будущую модель, а также те признаки, которые окажутся шумовыми.

Методика измерения качества и критерий успеха

Т.к. рассматривается задача бинарной классификации, то в качестве основной метрики измерения качества различных классификаторов была выбрана площадь под ROC-кривой (AUC ROC.). Эта метрика достаточно устойчива к дисбалансу классов (в задаче ощутим сильный дисбаланс: объектов класса «отток» только около 7%, тогда как класса «не отток» - около 93%), поэтому в дальнейшем при построении модели классы балансировались (методом Undersampling). Стоит напомнить, что **чем выше показатель AUC ROC, тем выше качество классификатора.** В то

же время хорошим показателем качества предсказания может выступать **точность** или **коэффициент неправильной классификации**: то есть, какое количество склонных к оттоку клиентов, предсказанных классификатором, на самом деле не ушли (TPR).

Дополнительными метриками качества могут служить полнота (recall), показывающая сколько от общего числа реальных пользователей, склонных к оттоку, было предсказано, как класс «отток», и f-мера.

Для тестирования качества модели от общего датасета была отделена выборка в 10 тыс. последних объектов, которые в дальнейшем не участвовали в обучении классификатора и использовалась только для контроля качества решения. Такая выборка позволяет убедиться, что при обучении модели не произошло переобучении либо какие-то другие ошибки.

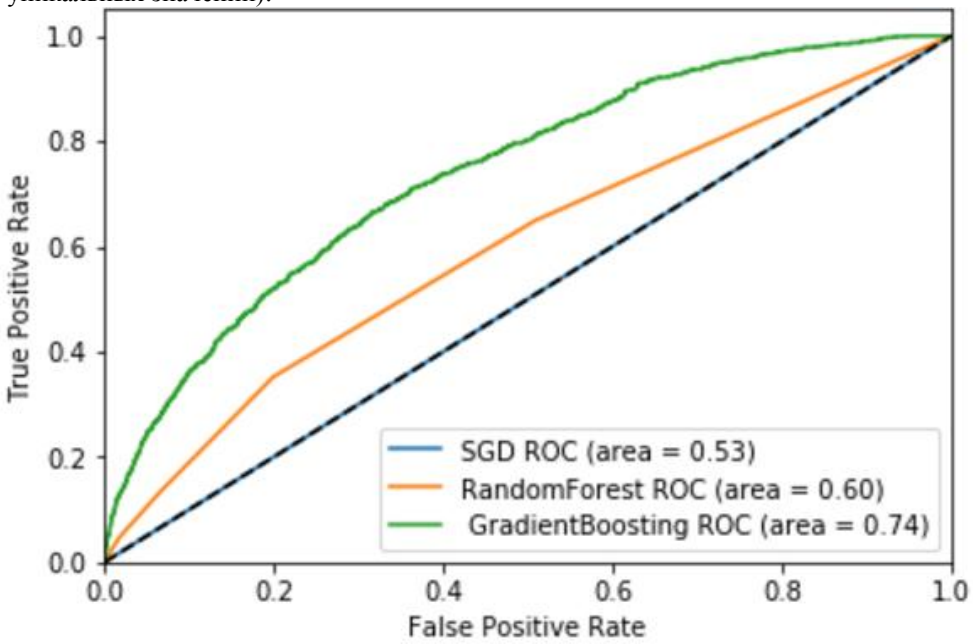
Для контроля качества модели в процессе обучения использовалась **кросс-валидация**. **Стратегией кросс-валидации** была выбрана stratified k-fold на 10 фолдов, т.к. она сохраняет баланс классов при этом все наблюдения используются и для тренировки, и для тестирования модели, и каждое наблюдение используется для тестирования в точности один раз.

Критерием успешности модели можно считать значение AUC_ROC больше, чем 0.68 – уровень baseline-решения в соревновании на Kaggle.

Техническое описание решения.

I. Предобработка данных и построение baseline-решения.

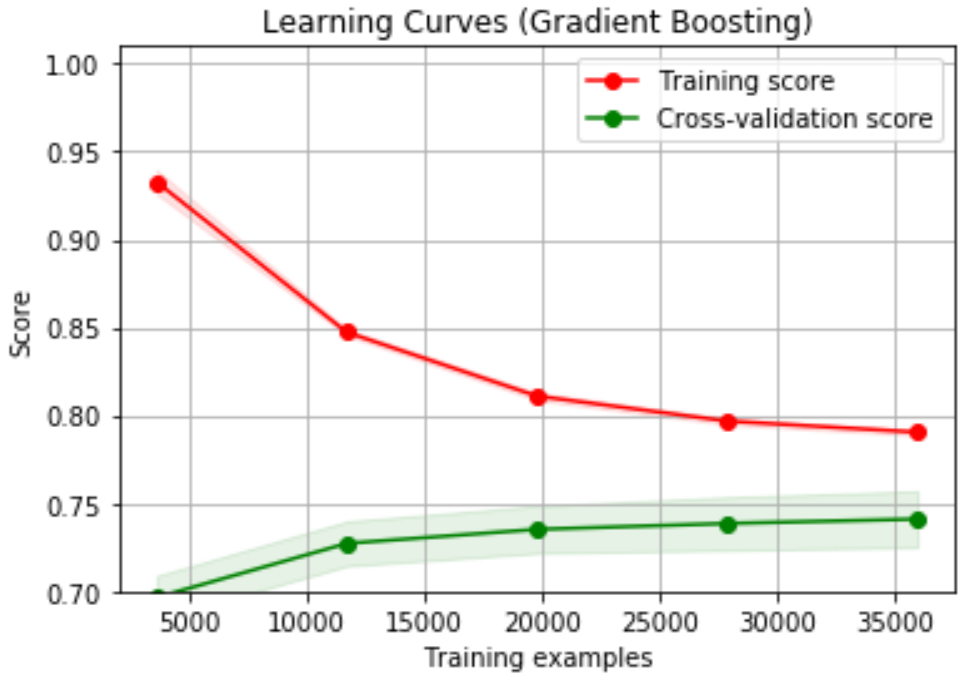
Для начала, **данные были обработаны**. Датасет был разделен на два датасета – один включал в себя только вещественные переменные, другой – только категориальные (столбец с метками классов оставили без изменения). Затем из обоих датасетов удалили признаки, полностью состоящие из пропущенных значений. Далее, пропущенные значения вещественного датасета заполнили средним по столбцам. Пропущенные значения категориальных признаков привели к строковому виду. Все категориальные переменные закодировали, заменяя каждую категорию числом входящих в неё объектов. Выбираем данный метод, т.к. в отличие от pandas.get_dummies или OneHotEncoder, данный метод не предусматривает значительного увеличения количества признаков (а в нашем случае есть переменные, у которых более 10800 уникальных значений).



После преобразований, были выбраны и построены три baseline-модели: SGDClassifier, случайный лес на 10 деревьев и градиентный бустинг. По всем моделям были посчитаны метрики качества на кросс-валидации – ROC_AUC, точность, полнота и f-мера. По лучшим результатам основной метрики – ROC_AUC, был выбран лучший классификатор – градиентный бустинг, который показал ROC_AUC 0.73, точность – 0.67, полноту – 0.5, f-меру – 0.49. Соответственно, в качестве baseline-решения был **выбран градиентный бустинг**.

II. Pipeline обработки данных.

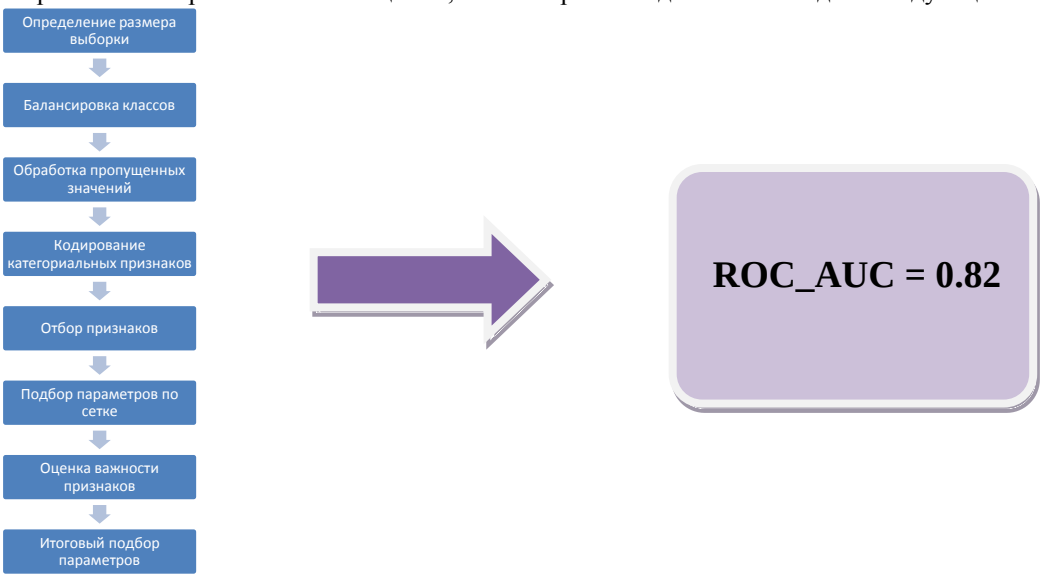
Теперь данные обрабатывались более тщательно. Для начала построили **кривые обучения**, чтобы определить оптимальный размер выборки, т.к. с ростом количества объектов обучающей выборки с определенного момента качество модели перестает расти. В нашем примере качество модели на кросс-валидации растет логарифмически с увеличением количества обучающих объектов примерно до 0.72-0.73 (после 20 тыс. объектов), а затем значительное увеличение качества не происходит.



Далее происходила балансировка классов методом undersampling, пока доли классов не сравнялись. После балансировки классов, пропущенные вещественные значения были заменены на средние по столбцам, а категориальные – приведены к строковому типу.

Далее категориальные данные были закодированы кодировщиком DictVectorizer.. Затем была применены стратегия отбора признаков на основе решающего дерева, после чего производился подбор параметров по сетке. Лучшим классификатором на данном этапе оказался градиентный бустинг на 40 деревьев, значение ROC_AUC с подбором параметров на обучении составило 0.7896, тесте - 0.74.

Далее с помощью метода «feature_importances» нашего классификатора, была оценена важность признаков, и в модель были отобраны только те признаки, важность которых была больше 0. В этом случае метрики ROC_AUC (после подбора параметров) были равны: на обучении - 0.8157, на тесте - 0.7452. Таким образом, качество модели по сравнению с baseline-решением выросло на 0.09. В целом, схема обработки данных выглядела следующим образом:



Качество модели и экономический эффект.

На каждом этапе обработки данных происходила оценка качества модели на кросс-валидации по метрике ROC_AUC. В итоге, лучшим результатом стало **значение данной метрики 0.82** - это значит, что в 82% случаев модель предсказывает большую вероятность ухода пользователей, действительно склонных к оттоку. При этом **точность составила 74%,** что означает, что в 74% случаев пользователи, склонные к оттоку по прогнозу модели, действительно уходят. Важнее правильно предсказать именно отток, т.к. если пользователь не собирается уходить, а его отнесут к классу «отток», то в этом случае убыток составит только расходы на кампанию по удержанию, тогда как если клиент собирается уходить, а модель отнесет его к классу «не отток», то компания не сможет вовремя предпринять меры по удержанию и лишится дополнительной прибыли от данного клиента.

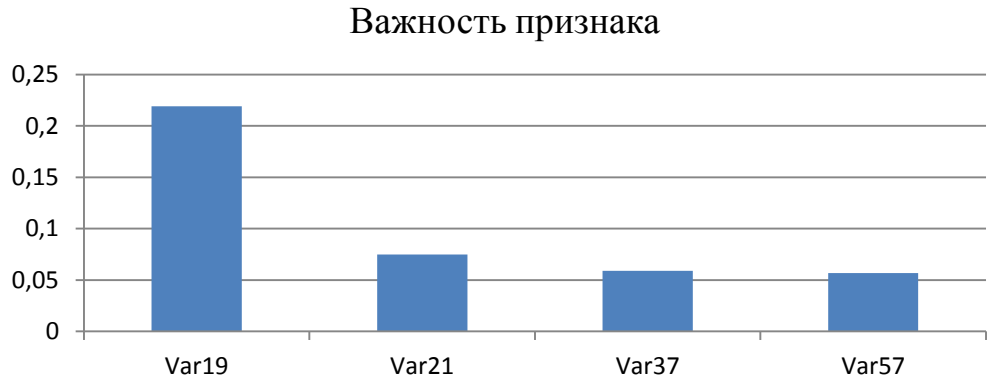
В целом, ROC AUC построенного классификатора превосходит baseline-решение на ресурсе Kaggle, что можно считать хорошим результатом.

Благодаря построенному классификатору можно оценить предполагаемый **экономический эффект** от проведения кампании по удержанию клиентов. Логично предположить, что экономический эффект будет максимальным, если полностью предотвратить отток пользователей (чем выше точность модели, тем выше эффект), однако, в этом случае и расходы на удержание также будут максимальными. Однако, благодаря предсказанию вероятности оттока, можно **оценить оптимальную долю пользователей**, склонных к оттоку, для которых проводить данную кампанию. Затем эту долю можно контролировать и менять при изменении других параметров модели, считающей экономический эффект. В то же время, при росте расходов на удержание одного клиента, необходимо, чтобы качество модели также росло.

Итог работы.

В результате выполнения задачи по прогнозированию оттока пользователей телекоммуникационной компании, был построен бинарный классификатор, предсказывающий вероятность отнесения клиента к классу «отток» с точностью 0.74 и метрикой ROC_AUC 0.82. В результате применения на практике полученного классификатора, можно рассчитать оптимальную долю пользователей, склонных к оттоку, тем самым увеличив прибыль компании при оптимальных расходах на кампанию по удержанию клиентов, что приведет к положительному экономическому эффекту для компании.

Кроме того, в ходе построения модели были выделены **наиболее важные признаки.** Это поможет для интерпретирования модели, а также позволит нам сформулировать новые гипотезы о том, по каким причинам пользователи уходят, с какими проблемами они сталкиваются и каким образом на это можно повлиять.



Для оценки экономического эффекта от использования нашей модели в продакшн, можно провести А/Б-тестирование на небольшой группе пользователей.

Дизайн А/Б тестирования.

Для начала, пользователи делятся на две группы: «А» и «Б». В первой группе, «А», никаких изменений происходить не будет и по ней в дальнейшем будем оценивать эффективность использование нашей модели. Для пользователей группы «Б» будем **прогнозировать отток клиентов** с помощью нашей модели и **рассчитывать экономический эффект от их удержания (т.е. прибыль)**.

Применим полученный классификатор к пользователям из группы «Б», на выходе получим список вероятностей оттока. Для оптимальной доли пользователей, предсказанных моделью как класс «отток», проведем маркетинговую кампанию по удержанию (это могут быть скидки, подарки, улучшение сервиса и т.д.).

Через определенный промежуток времени (например, три месяца, т.к. разумно считать среднемесячный показатель оттока) останавливаем кампанию и считаем прибыль от удержанных клиентов. Считаем разницу между показателями оттока группы «А» и группы «Б», считаем разницу в прибыли от удержания клиентов группы «Б» и затратах на проведение программы по удержанию. Оцениваем эффективность использования модели.

Улучшение и дообучение модели.

В ходе анализа экономического эффекта от внедрения нашего классификатора, был сделан вывод, что улучшение качества модели позитивно сказывается на экономическом эффекте. Для дальнейшего улучшения качества модели следует тщательно проанализировать объекты, на которых наша модель ошибается чаще всего (предсказывает неверный класс с большой вероятностью). Для начала можно посмотреть на распределение ошибок и понять от чего оно зависит. Например, можно выбрать несколько наиболее важных признаков и построить гистограммы распределения ошибок на данных признаках. Кроме того, следует оценить инвестиции в исправление данных ошибок и прибыль от их исправления. С течением времени модель начнет устаревать. Важно определить этот момент, для чего нужно постоянно оценивать метрики качества и следить за данными для обучения (за появлением новых признаков, за изменением поведения пользователей).