Прогнозирование оттока клиентов

Цели и задачи проекта:

Задача прогнозирования оттока является одной из важнейших подзадач в области работы с аудиторий и актуальна не только для телекоммуникационных компаний, но и для большинства организаций, оказывающих услуги в сегменте B2C (прим. часто и в B2B тоже, однако в этом случае под клиентом мы понимаем компанию). Такие задачи часто возникают на практике у телекоммуникационных операторов, провайдеров кабельного телевидения, страховых компаний, банков, крупных и средних интернет-порталов и др. Решение задачи по удержанию клиентов является более предпочтительным в сравнении с привлечением новых клиентов, связанным с большими затратами.

Цель проекта - научиться находить пользователей телеком компании, склонных к оттоку. Если научиться находить таких пользователей с достаточной точностью заблаговременно, то можно эффективно управлять оттоком: например, выявлять причины оттока; помогать пользователям, попавшим в группу риска, решать их проблемы и задачи; проводить кампании по удержанию.

В терминах машинного обучения для решения задачи прогнозирования оттока клиентов строятся вероятностные модели бинарной классификации, где целевой класс представляют собой пользователи, покидающие сервис – отток.

Описательный анализ данных:

На этапе анализа данных был обнаружен дисбаланс классов: класс “отток” – 7.44%, класс “не отток” – 92.56%.

Анализ показал наличие большого количества пропусков: 154 признака из 230 имеют 90% и более пропусков (количество непрерывных признаков с более чем 90% пропусков – 148, количество категориальных признаков с более чем 90% пропусков – 6).

Распределения в разрезе классов наиболее коррелированных числовых признаков с целевой переменной показало наличие выбросов и разряженность данных (рис. 1).

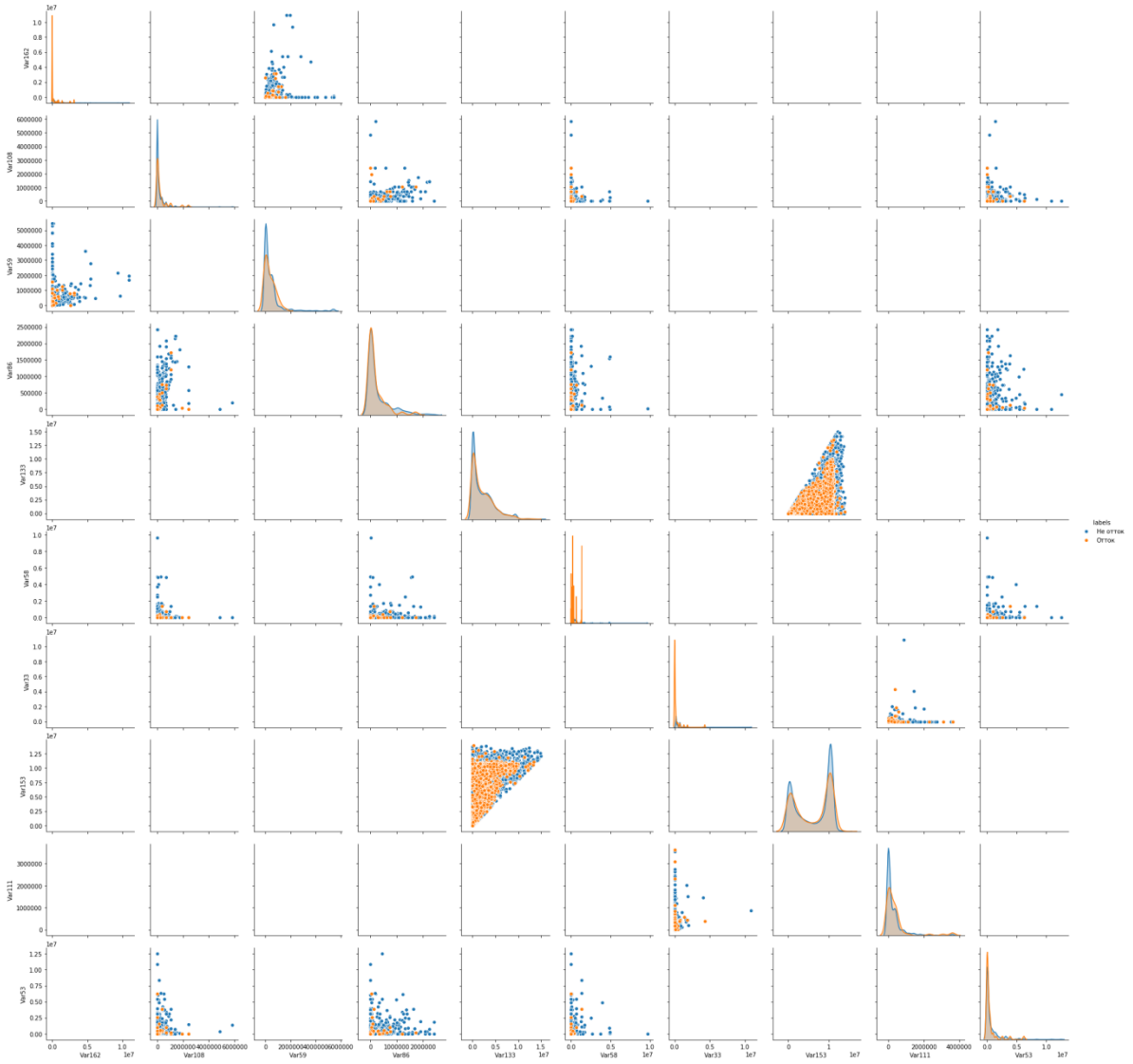


рис. 1

Распределения в разрезе классов наименее коррелированных числовых признаков с целевой переменной показало наличие выбросов, разряженность данных, а также то, что некоторые числовые признаки дискретны (рис. 2).

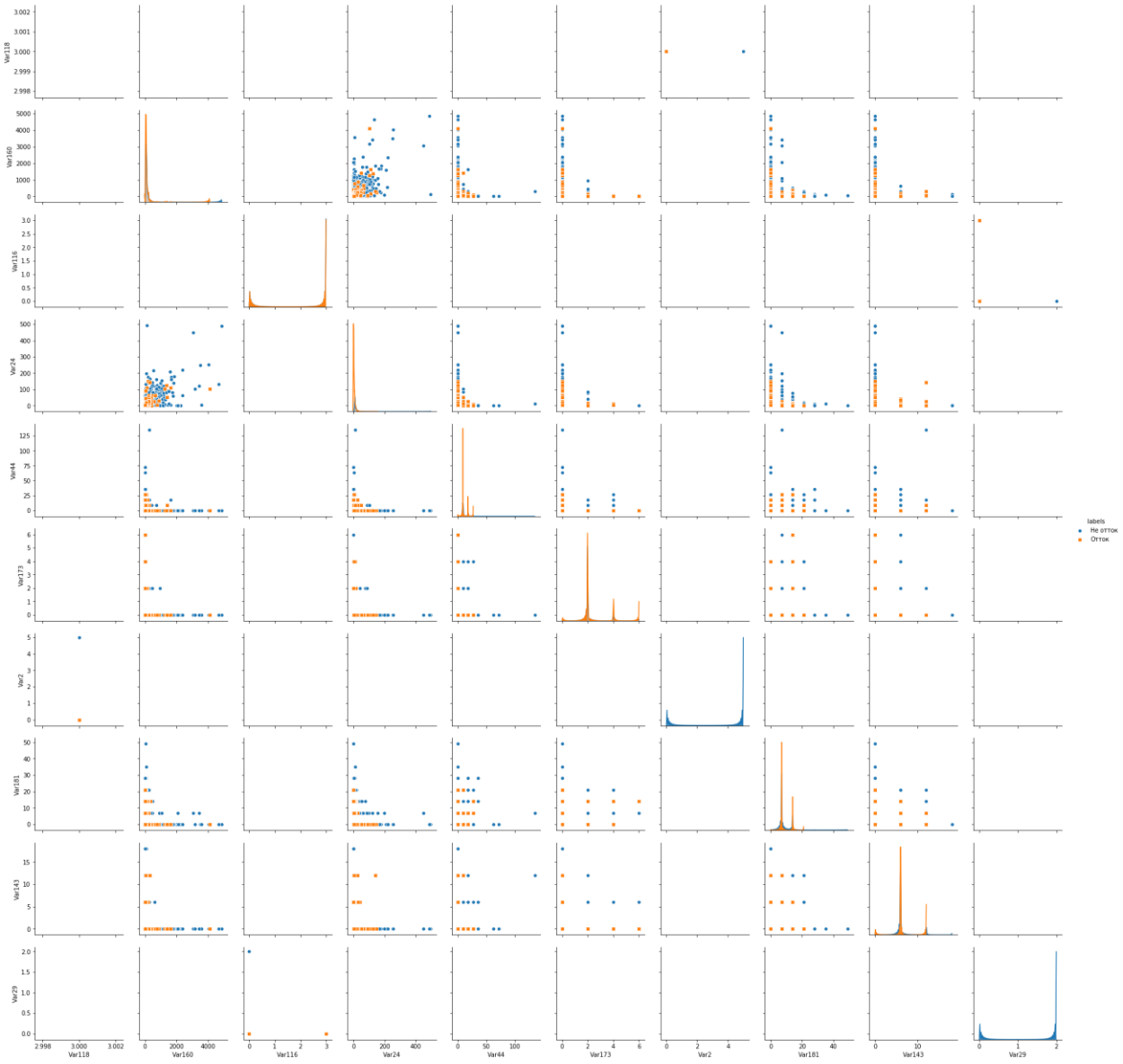


рис. 2

11 категориальных признаков имеют более 100 уникальных значений, 9 из них имеют несколько тысяч уникальных значений, что затрудняет их обработку.

Методика измерения качества и критерий успеха.

В условиях несбалансированности классов лучше оценивать с помощью метрики AUC-ROC. ROC-кривая строится в осях FPR и TPR, которые нормируются на размеры классов. Следовательно, при изменении баланса классов величина AUC-ROC и неизменных свойствах объектов выборки площадь под ROC-кривой не изменится. Значение AUC-ROC имеет смысл вероятности того, что если были выбраны случайный положительный и случайный отрицательный объекты выборки, положительный объект получит оценку принадлежности выше, чем отрицательный объект.

Метрики F1-мера, precision, recall не стоит использовать, поскольку они чувствительны к дисбалансу классов.

Также использован коэффициент Мэттьюса, вычисленный по перекрёстной таблице по истинным и предсказанным меткам (TP, FP, TN, FN).

Тестирование лучше проводить на отложенных данных – 25% от исходного датасета, возможно AB – тестирование. При тестировании следует ориентироваться на экономический эффект от внедрения модели – это покажет реальный результат работы, а также влияние ошибок первого и второго рода. Критерием успешности можно считать увеличение площади под ROC – кривой (отталкиваясь от случайного угадывания = 0.5), увеличение экономического эффекта на этапе тестирования на отложенном датасете (или на AB - тесте).

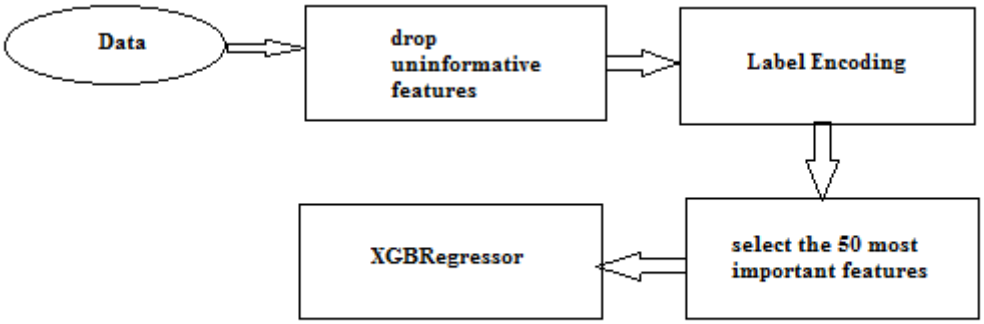
Техническое описание решения.

В ходе эксперимента были опробованы следующие подходы к построению модели:

1) Задание весов объектам выборки.
2) Балансировка классов: under-/oversampling.
3) Заполнение пропусков в данных:
    - заполнение пропусков в численных признаках нулями, в категориальных - наиболее часто встречающимися значениями;
    - заполнение пропусков в численных признаках медианными значениями, в категориальных - наиболее часто встречающимися значениями;
    - заполнение пропусков в численных признаках средними значениями, в категориальных - наиболее часто встречающимися значениями;
    - замена пропусков наиболее часто встречающимся значением в признаке;
    - замена пропусков в числовых признаках средним, если количество уникальных значений больше 180, иначе наиболее часто встречающимся значением в признаке; в категориальных - замена наиболее часто встречающимся значением в признаке;
    - заполнение пропусков нулями.
4) Обработка признаков:
    - обработка категориальных признаков с помощью OrdinalEncoder;
    - обработка категориальных признаков с помощью LabelEncoder;
    - Дискретизация численных признаков с числом уникальных значений более или равным 100, дальнейшее применение ordinalEncoding. Для численных признаков с числом уникальных значений менее 100, применение labelEncoding. labelEncoding для категориальных признаков.
    - Дискретизация численных признаков с числом уникальных значений более или равным 100, дальнейшее применение oneHotEncoding. Для численных признаков с числом уникальных значений менее 100, применение labelEncoding. labelEncoding для категориальных признаков.
5) Lasso, Ridge – регуляризация; отбор признаков на основе их важности, определённой моделью.
6) Классификаторы: RidgeClassifier, LoghisticRegression, RandomForestClassifier, GradientBoostingClassifier. Наилучшее качество показал градиентный бустинг.
7) Поиск оптимальных гиперпараметров XGBoostRegressor с помощью GridSearchCV.

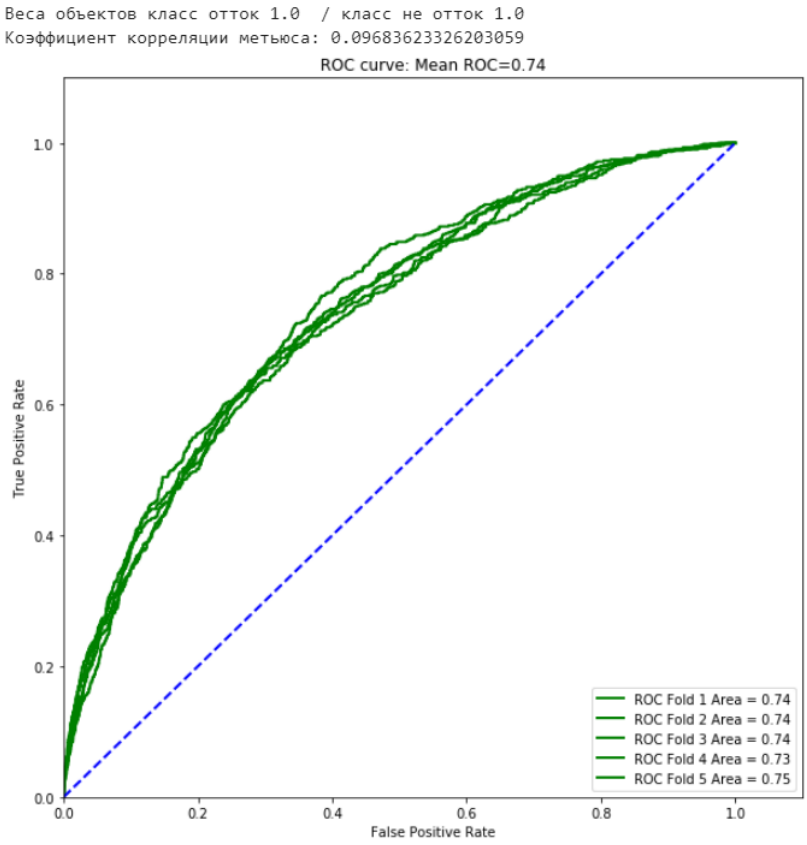
На основе проведённых экспериментов определён следующий pipeline, показавший лучшее качество. Он был использован для построения модели.

1) Заполнение пропусков нулями.
2) Отбрасывание неинформативных признаков (константные признаки).
3) LabelEncoding – для категориальных признаков.
4) Отбор 50 самых важных признаков.
5) Обучение XGBoostRegressor с оптимальными гиперпараметрами.



Оценка качества модели:

Модель оценивалась по метрике AUC-ROC. Итоговая оценка осуществлялась по отложенной выборке. Кросс-валидация на обучении на 5 фолдах.



На отложенной выборке метрика ROC-AUC составила 0.736.

Данная модель построена без undersampling’a. Хотя при использовании undersampling’a коффициента мэттьюса больше, на экономическом эффекте это не сказывается. При использовании undersampling’a уменьшается количество ошибок первого рода, но увеличивается количество ошибок второго рода.

Следующие признаки ввели наибольший вклад в модель:

feature_importances	feature
0.047488	Var126
0.036449	Var218
0.033596	Var207
0.032633	Var229
0.028338	Var73
0.028057	Var205
0.026993	Var74
0.018390	Var189
0.016287	Var217

После построения модели был оценён экономический эффект. Для этого были введены следующие параметры:

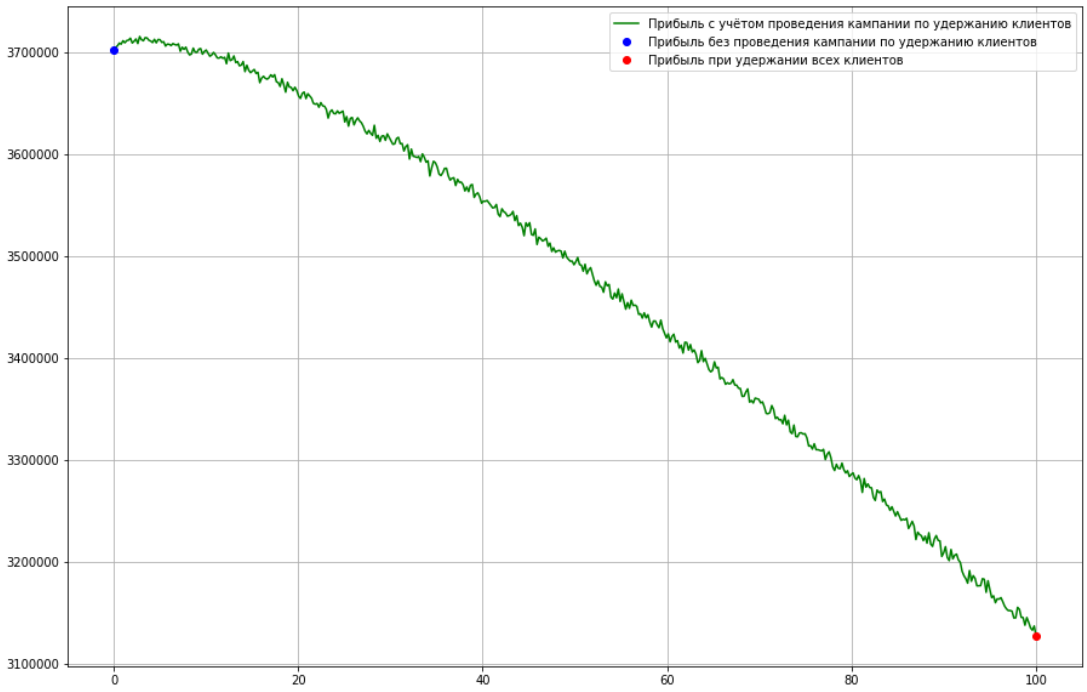
- сколько денег в среднем приносит один пользователь в месяц;
- сколько денег в среднем вы будете вкладывать в удержание одного пользователя;
- с какой вероятностью пользователь примет ваше предложение;
- сколько пользователей (например, топ 1% или топ 25% согласно ранжированию по вашей модели) будет участвовать в кампании.

В связи с недоступностью информации о доходе компании с одного человека и других параметрах, использованы искусственные значения показателей.

Параметры:

- в среднем приносит один пользователь в месяц 400р.
- сколько денег в среднем вы будете вкладывать в удержание одного пользователя:
- 15% скидка пользователям, склонным к оттоку на срок 12 месяцев;
- затраты на работу сотрудников (телефонный разговор): среднее время разговора 8 минут, зарплата сотрудника 30т.руб. зарплата сотрудника/ 22 рабочих дня/ 8 часов/ (8/60) = 22.7 рубля;
- с вероятностью 70% пользователь примет предложение;
- топ ранжированию по модели будет участвовать в кампании по удержанию.

Ниже представлена кривая прибыли с учётом проведения кампании по удержанию топа клиентов в процентах (прибыль от 10000 клиентов из отложенной выборки).



При удержании всех клиентов прибыль компании составит 2510200.

Прибыль компании без проведения мер по удержанию клиентов составит 3700000.

Прибыль компании при удержании оптимального топа клиентов составит около 3716000.

Оптимальный размер топа клиентов для удержания составляет 4-5 процентов.

Выводы:

Для данного эксперимента оптимальный размер топа составил 4-5 процентов. Полученная модель показывает хороший экономический эффект. Минимальную прибыль показал подход при полном удержании.

Для увеличения экономического эффекта нужно уменьшать количество ошибок первого и второго рода.

Вложение средств в улучшение качества модели может быть оправданным в том случае, если это повлечёт снижение количества ошибок первого и второго рода. Чтобы оценить на сколько нужно улучшить модель, чтобы это качественно сказалось на экономическом эффекте от удержания, нужно оценить затраты на увеличение качества модели, а также предполагаемое (желаемое) изменение экономического эффекта от увеличения качества модели.

Необходимо проводить мониторинг за качеством модели. Возможно, придётся её дообучать или перестраивать при появлении новых услуг, изменении предпочтений пользователей другим услугам, при появлении новых компаний (например, при предоставлении пакетов услуг) на телеком рынке.